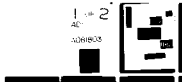
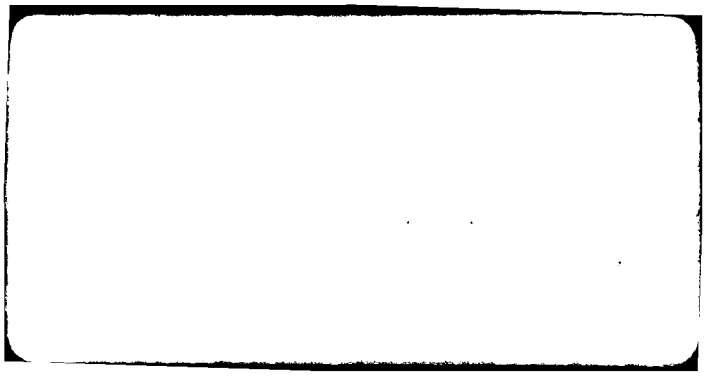
AD-A081 803 MAXIMUS INC MCLEAN VA F/G 12/1
FURTHER RESEARCH INTO A NON-PARAMETRIC STATISTICAL SCREENING SY--ETC(U)
DEC 79

UNCLASSIFIED NL





MAXIMUS

Prepared for:

Office of Naval Research
Department of the Navy

(6) FURTHER RESEARCH INTO A NON-PARAMETRIC STATISTICAL SCREENING SYSTEM,

December 14, 1979

(11) 14 Dec 79 (12) 104 1

Funds for this project were supported by the Statistics and Probability Program, Office of Naval Research under Contract NR 042-401.

MAXIMUS, Inc.
6723 Whittier Avenue
McLean, Virginia 22101
(703)734-0050

ABSTRACT

A new statistical technique is introduced for screening a population on the basis of their observed characteristics. The technique treats nominal independent variables with a binary dependent variable. Different objective functions are specified for constructing different decision rules. A recommended decision rule results from achieving a proportionate reduction in error (PRE), by using information in the independent variables rather than just the dependent variables. Current approaches to screening are compared using five desirable properties postulated for decision rules. A Monte Carlo simulation approach is used to construct decision rules using Boolean operators. Finally, a General Sequential Algorithm is presented.

TABLE OF CONTENTS

TABLE OF CONTENTS

(Continued)

I. INTRODUCTION

I. INTRODUCTION

1.1 Statistical Screening

The use of statistical techniques to distinguish members from two or more population groups got its greatest impetus from the development of the linear discriminant function by Fisher in 1935. Although many advancements in the state-of-the-art have been made since that time, the LDF remains the most prevalent technique.

Among the many applications of statistical screening to practical problems, the following are typical:

- screening tax returns for cases with underestimated tax liability;
- identifying college football players with the greatest potential for success in the NFL;
- identifying potential fraud in public assistance programs;
- screening potential borrowers for credit-worthiness;
- screening suspected criminals for prosecution;
- identifying disease prone individuals.

In general, the aim is to identify two groups so that members of each group may be treated differently.

Similar applications may occur for the Navy. In particular, consider the following:

- aviation officer program attrition has been a major problem. Costs associated with in-flight training attrition have been estimated at over $40 million per year. A 10% decrease in attrition in each phase of flight training could result in a savings of over $2 million per year;
- the attrition rate among first-term enlistees is 10% during recruit training and another 7% in the remainder of the first year. At

an estimated cost in excess of $5,000 per first-year failure, a 25% reduction in attrition would result in savings in excess of $20 million.

Thus, the potential for improving the screening process for recruits or trainees is very high if an effective technique can be developed. In the next section, we review briefly some of the important considerations in screening.

1.2 Considerations in Screening

In this section, we introduce three of the major considerations relevant to a screening problem:

- the descriptive variables;
- the interaction among variables;
- the objective function.

1.2.1 Variables

In most applications, the researcher has a set of variables $X_1, \ldots, X_k$ which can be used to describe population members. The intent is to use these variables to distinguish members from each group. The variables themselves may be categorized on the basis of their scale of measurement:

- a nominal scale is used to distinguish between different classes. The particular values taken on by a nominal scale variable have no particular significance. They may be renamed, relabeled or reordered without changing the meaning;
- an ordinal scale is used if there is an underlying ordering of classes;
- an interval scale is used if the difference between two values has meaning;
- a ratio scale is an interval scale with a true zero point.

The scale definitions are such that each scale incorporates the properties of the preceding scale. Thus, a variable measured on a ratio scale provides more information than a variable on a nominal scale, for example. On the other hand,

mathematical operations that may be meaningful for an interval or ratio scale might not be applicable to variables measured on a nominal or ordinal scale.

Nonetheless, statistical screening techniques are often applied without attention to the types of variables describing population members. In particular, users apply sophisticated mathematical procedures to variables measured at the nominal level. Some of the problems encountered with such techniques are described in Chapter III. One of the major objectives of this research, then, is to develop a class of statistical screening techniques which are designed to be compatible with, and meaningful for, nominal level variables.

1.2.2 Interaction Among Variables

In many practical applications, especially with human populations, it may be the combination of variable values describing an individual that is a distinguishing factor. Some statistical techniques focus on the significance of single variables, considered one at a time, and miss important relationships that exist (see Chapter II). With the increased power of computers, the ability to detect interaction effects has been greatly enhanced. The approaches we develop here take advantage of this power.

1.2.3 Objective Functions

Much of the statistical research pertaining to statistical screening has focused on the probability of misclassification as the appropriate function to be minimized. In practice, however, the ultimate user may have a different objective function related, perhaps, to cost, resources and other constraints. In Chapter II, we discuss alternative objective functions of interest and in Chapter III we state that a desirable property of a statistical screening procedure is that it be flexible with respect to the types of objective functions that can be handled. Again, a major objective of this research is to develop a class of statistical screening procedures having this property.

1.3 Research Questions

With the above considerations in mind, then, the following research questions may be identified:

- Can the screening problem be characterized in a uniform fashion? What objective functions are pertinent to the screening problem? How are they related?

- How can alternative techniques be compared on a statistical basis? Can confidence intervals and hypothesis testing procedures be developed for parameters of interest?

- What properties should a screening technique have? How do some of the well-known screening techniques fare with respect to these properties?

- What new procedures can be developed to handle the type of problem under consideration? How well have these procedures performed in actual practice? What are the features of these procedures?

The ultimate objective of the research is to develop a new class of statistical screening procedures that have general applicability to a wide range of practical problems.

1.4 Overview of the Report

While much of statistical research must, by necessity, deal with theoretical considerations with, perhaps, restrictive or unrealistic assumptions, there is also a strong need for applications oriented research that may open up the field of statistics to a wider range of actual problems. The research in this report, while based on sound statistical underpinnings, is definitely oriented towards practical applications. This is both a strength and a weakness. It is a strength in that there are many real-life situations in which the results could immediately be put to use; it is a weakness in that the results are based, to a large extent, on intuition and experience rather than on rigorous mathematical development. The nature of the problem makes it rather intractable for closed-form analysis.

Nonetheless, the research presented here represents a new step in the field of statistical screening, a step that we believe contributes greatly to the current state-of-the-art. The closest parallel to the type of research presented here is the work of Sonquist and Morgan in developing the Automatic Interaction Detection (AID) technique. The rapid popularity AID has gained as a technique in the few years since it was developed is a testament to the need for applications-oriented research in this field.

The report is organized as follows:

- In Chapter II, we define the problem framework and underlying assumptions. We show how the screening problem can be characterized by three parameters. We present some of the potentially relevant objective functions and use some results from the field of prediction logic.

- In Chapter III, we review the major competing techniques in terms of five properties that we believe should be held by an effective screening technique.

- In Chapter IV, we present an initial approach taken to development of a technique satisfying the above techniques. Results of an empirical test of the approach are presented.

- In Chapter V, we develop a class of statistical techniques, based upon a General Sequential Algorithm.

II. PROBLEM FRAMEWORK

II. PROBLEM FRAMEWORK

2.1 Introduction

In this chapter, we present the basic framework for characterizing the screening problem. First, we introduce the assumptions and notation for the type of screening situation under consideration. Then, we demonstrate that the problem can be specified in terms of three basic parameters. This leads to a discussion of possible objective functions for the decision problem. Then, we relate the screening problem to the field of prediction logic and, using this relationship, develop some results that are useful for defining and evaluating procedures.

2.2 Notation and Assumptions

Assume that we have a random sample of size n from a mixed population $\Pi=\Pi_1 \cup \Pi_2$, where $\Pi_1 \cap \Pi_2=\phi$. Each observation is a vector $\vec{x}$ from the sample space $\mathbb{X}=(X_1, \ldots, X_p)$ where each X_i, $i=1, \ldots, p$, can take on any of s_i discrete values. We also assume that the population membership of each sample observation is known.

The aim of a statistical screening procedure is to develop decision rules $D = \langle D_1, D_2 \rangle$, $D \varepsilon \mathcal{D}$, of the type:

if $\vec{x} \varepsilon D_1$, we assign $\vec{x}$ to Π_1

if $\vec{x} \varepsilon D_2$, we assign $\vec{x}$ to Π_2.

That is D_1 and D_2 are partitions of the sample space $\mathbb{X}$. We are restricting the analysis to the case where $D_1 \cup D_2=\mathbb{X}$ and $D_1 \cap D_2=\phi$. That is, every observation is assigned to Π_1 or Π_2. Some researchers allow the possibility of a set $D_3=(D_1 \cup D_2)^c$ in which no decision is made or new observations are assigned to Π_1 or Π_2 according to some random process.

Thus, given a decision rule $D= \langle D_1, D_2 \rangle$, the population Π may be partitioned into a 2 x 2 table as shown below:

	D_1	D_2	
Π_1	P_{11}	P_{12}	$P_{1\cdot}$
Π_2	P_{21}	P_{22}	$P_{2\cdot}$
	$P_{\cdot 1}$	$P_{\cdot 2}$	1

where the $P_{ij} = P(\Pi_i, D_j)$, $i,j = 1,2$. For the sample, the 2 x 2 table contains the sample frequencies p_{ij}.

To simplify the notation, consider a reduced set of parameters. Specifically, let

$$P(\vec{x} \epsilon D_1) = P(D_1) = P$$
$$P(\vec{x} \epsilon \Pi_1 | \vec{x} \epsilon D_1) = P(\Pi_1 | D_1) = Q$$
$$P(\vec{x} \epsilon \Pi_1) = P(\Pi_1) = E = 1 - P(\Pi_2)$$

Then, the table entries are as follows:

$$P_{11} = P(D_1, \Pi_1) = PQ$$
$$P_{21} = P(D_1) - P(D_1, \Pi_1) = P - PQ = P.(1-Q)$$
$$P_{12} = P(D_2, \Pi_1) = P(\Pi_1) - P(D_1, \Pi_1) = E - PQ$$
$$P_{22} = P(\Pi_2) - P(D_2, \Pi_1) = 1 - P - (E - PQ).$$

The entries are shown below:

	D_1	D_2	
Π_1	PQ	E-PQ	E
Π_2	P.(1-Q)	1-P-(E-PQ)	1-E
	P	1-P	1

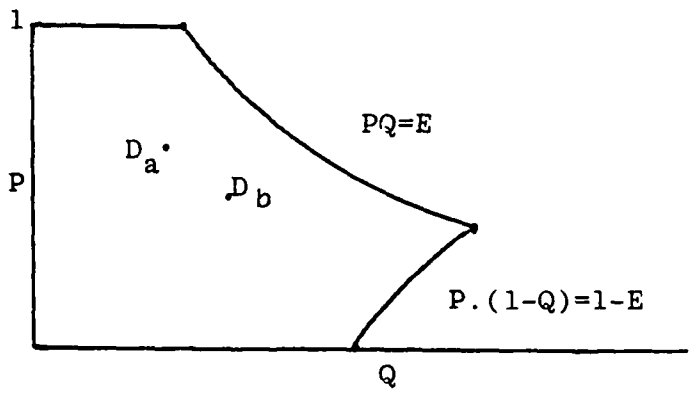
Since E is a constant, decision rules D are characterized by the pair (P, Q).

The corresponding statistics for any sample are given by p, q and e. Note that e is a constant for any given sample being used to construct decision rules D.

It is possible to construct a boundary that contains all decision rules by considering the following constraints:

1. $0 \leq P, Q \leq 1$ (since P and Q are probabilities).
2. $PQ \leq E$ $\left(\text{since } E = P(D_1, \Pi_1) + P(D_2, \Pi_1) \geq P(D_1, \Pi_1) = PQ\right)$
3. $P.(1-Q) \leq 1-E$ $\left(\text{since } 1-E = P(D_1, \Pi_2) + P(D_2, \Pi_2) \geq P(D_1, \Pi_2) = P.(1-Q)\right)$.

Exhibit 2.1 illustrates the form of this boundary:



Note that when PQ=E, $\Pi_1 \subset D_1$. That is, D_1 contains all of Π. When P.(1-Q)=1-E, D_1 contains all of Π_2.

The key question that arises from this description, however, is the following. Given two decision rules D_a and D_b, as depicted in the exhibit, which one is better? That is, before we can develop procedures for developing decision rules, we must have some concept of what a "good" decision rule is. In the material that follows, we develop some theory of objective functions to answer this question.

2.3 Objective Functions

2.3.1 Minimize Probability of Misclassification

The most prevalent objective function used in screening problems is the probability of misclassification. For a decision rule $D = <D_1, D_2>$, the true probability of misclassification is

$$t(D) = P(D_1, \Pi_2) + P(D_2, \Pi_1), \text{ or}$$

$$t(D) = \underset{D_1}{\Sigma} L_2(\vec{x})\,(1-E) + \underset{D_2}{\Sigma} L_1(\vec{x})\,E$$

where $L_i(\vec{x})$ is the density of $\vec{x}$ under population Π_i, i = 1,2.

In terms of the parameters defined in 2.2, we have

$$t(D) = P.(1-Q) + E - PQ$$

$$= E + P.(1-2Q) \qquad (1)$$

Since the decision rule is based on a sample from Π, the estimated probability of misclassification is

$$\hat{t}(D) = e + p\,(1-2q). \qquad (2)$$

Therefore, a reasonable objective for the screening process is to find a decision rule $D^* = < D_1^*, D_2^* >$ to minimize $t(D)$. That is,

$$t(D^*) = \inf_{D \in \mathcal{D}} t(D) = \text{optimum probability of misclassification}$$

where $\mathcal{D}$ is the class of all possible decision rules.

In practice, we are dealing with a sample and the corresponding objective is to minimize $\hat{t}(D)$, that is, to find D^{**} such that

$$\hat{t}(D^{**}) = \inf_{D \in \mathcal{D}} \hat{t}(D) = \text{estimated optimum probability of misclassification.}$$

Several authors [Cochran and Hopkins (1961); Mills (1966); Mickey (1968); Glick (1972, 73) and Goldstein and Wolf (1977)] have studied the relationships among $\hat{t}(D^{**})$, $t(D^*)$ and $t(D^{**})$. They show that

$$E\left(\hat{t}(D^{**})\right) \leq t(D^*). \qquad (3)$$

That is, the estimated optimum misclassification probability tends to be an underestimate of the true optimum probability. This is intuitively reasonable since the screening procedure finds the best rule for the sample rather than the population. This is similar to the result in fitting regression models based on a sample: the sample R^2 tends to overestimate the true R^2.

Similarly, it holds that

$$t(D^*) \leq t(D^{**}), \qquad (4)$$

with equality only if $D^* \equiv D^{**}$. This is true by definition since

$$t(D^*) = \inf_{D \in \mathcal{D}} t(D) \leq t(D) \;\forall\; D \in \mathcal{D}.$$

Lackenbruch (1975) proposes a method to estimate $E\left(\hat{t}(D^{**})\right)$, called the mean apparent (misclassification) error. He uses n-1 points to classify the remaining point. This procedure is repeated for all points and he records the proportion misclassified for each group. His estimate is

$$E \cdot \frac{m_1}{n_1} + (1-E) \cdot \frac{m_2}{n_2} \qquad (5)$$

where n_1, n_2 are the number of sample cases from Π_1 and Π_2 respectively, and m_1, m_2 are the number misclassified from Π_1 and Π_2.

Note that Lackenbruch's estimate assumes knowledge of the true probabilities of membership in Π_1 and Π_2.

In fact, a jacknife procedure (Miller, 1974) may be used as follows: let $\hat{t}_{-j}$ denote the probability of misclassification with the jth observation removed. Then, the jacknifed apparent classification error is

$$\hat{t}_J(D^{**}) = n\,\hat{t}(D^{**}) - \frac{n-1}{n} \sum_{j=1}^{n} \hat{t}_{-j}(D^{**}) \qquad (6)$$

Example of Bias

Consider the following trivial sample-based decision rule $D = \langle D_1, D_2 \rangle$ with

$$D_1 = \{\vec{x} \mid \vec{x}_s \in \Pi_1\}$$

$$D_2 = D_1^c$$

where $\vec{x}_s$ denotes the sample observations.

That is, a new observation is assigned to Π_1 if it matches exactly one of the sample observations from Π_1. Otherwise, it is assigned to Π_2. Assume further that there are no exact matches in the sample of observations that are in both Π_1 and Π_2, i.e., if $\vec{x}_s \in \Pi_1$ and $\vec{x}_t \in \Pi_2$, $\vec{x}_s \neq \vec{x}_t$. This assumption is quite reasonable for small samples with large k.

By construction,

$$\hat{t}(D) = 0 = \inf_{D \in D} \hat{t}(D) = \hat{t}(D^{**}) \leq t(D^{**})$$

That is, the sample probability of misclassification is zero. In many practical problems, this decision rule is available but is, of course, neither realistic nor useful.

2.3.2 Minimize Variance of Estimates

We can define indicator variables as follows:

$$Y_{\vec{x}} = \begin{cases} 1 \text{ if } \vec{x} \in \Pi_1 \\ 0 \text{ if } \vec{x} \in \Pi_2 \end{cases} \qquad \hat{Y}_{\vec{x}} = \begin{cases} 1 \text{ if } \vec{x} \in D_1 \\ 0 \text{ if } \vec{x} \in D_2 \end{cases}$$

where, as before, $D = \langle D_1, D_2 \rangle$ is any decision rule.

The estimated variance of the estimates, $\hat{\sigma}^2_{\hat{y}}$, is

$$\sum_{\vec{x}} \frac{(Y_{\vec{x}} - \hat{Y}_{\vec{x}})^2}{n} = \sum_{\vec{x} \in \Pi_1} \frac{(Y_{\vec{x}} - \hat{Y}_{\vec{x}})^2}{n} + \sum_{\vec{x} \in \Pi_2} \frac{(Y_{\vec{x}} - \hat{Y}_{\vec{x}})^2}{n}$$

$$= \sum_{\vec{x} \in \Pi_1} \frac{|Y_{\vec{x}} - \hat{Y}_{\vec{x}}|}{n} + \sum_{\vec{x} \in \Pi_2} \frac{|Y_{\vec{x}} - \hat{Y}_{\vec{x}}|}{n}$$

$$= P(D_2 \cap \Pi_1) + P(D_1 \cap \Pi_2)$$

= Estimated probability of misclassification = $\hat{t}(D)$.

Similarly, the true variance is $\sigma^2_{\hat{y}} = t(D)$.

Thus, we have shown that minimizing the probability of misclassification is equivalent to minimizing the variance of estimates.

2.3.3 Maximize R^2

The results of 2.3.2 can be extended to include a measure equivalent to the R^2 measure used in curve fitting:

$$R^2 = \frac{\text{Total Variance-Unexplained Variance}}{\text{Total Variance}} = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

$$\text{Total Variance} = \frac{\sum (Y - \bar{Y})^2}{n} = \frac{\sum Y^2 - n\,\bar{Y}^2}{n}$$

$$= \frac{n\,e - n\,e^2}{n} = e(1-e)$$

$$\text{Unexplained Variance} = \frac{\sum (Y_x - \hat{Y}_x)^2}{n} = e + p - 2\,pq$$

Thus,

$$R^2 = \frac{e\,(1-e) - (e+p-2pq)}{e(1-e)} = 1 - \frac{e+p-2pq}{e(1-e)} = 1 - \frac{\text{Probability of Misclassification}}{\text{Total Variance}}.$$

Since e is fixed for any sample, maximizing R^2 is equivalent to minimizing the probability of misclassification. Note also that

$$R^2 = 1 \iff e + p - 2pq = 0 \iff p = e \text{ and } q = 1$$

$$R^2 = 0 \iff e + p - 2pq = e\,(1 - e)$$

$$\iff p\,(1 - 2q) = -e^2.$$

2.3.4 Maximize Yield

In some practical applications, the aim may be to detect as many cases from Π_1 as possible. That is, the intent is to maximize PQ or, for the sample, to maximize pq. Note that pq/e is the sample proportion of Π_1 cases correctly classified into Π_1.

2.3.5 Maximize Q Subject to Fixed P

As discussed in Chapter I, the aim of the screening effort is to identify new cases that are likely to be from Π_1 so that they may be treated differently, e.g., auditing tax returns. However resources may be limited, for example, the IRS may only be able to audit 10% of each year's returns.

Thus, another objective function may be to maximize Q subject to fixed P = P*, say. Thus, the space of decision rules is reduced to those with a value of $P \geq P^*$.

When dealing with a sample, however, we only have the sample realizations p for each decision rule. For this situation, the following approach is suggested. Compute

$$p^* = P^* + Z_\alpha \sqrt{\frac{P^*\,(1 - P^*)}{n}},$$

where Z_α is the normal ordinate associated with a one-tailed probability of α.

Then, the sample decision rules are restricted to those with a realized $p \geqslant p^*$. Note that, for any decision rule,

$$P(P \geqslant P^* \mid p \geqslant p^*) \geqslant 1 - \alpha,$$

assuming the normal approximation to the binomial. An exact value for p* may also be computed if the requirements for the normal approximation do not hold.

Conversely, in some problems, it may be desirable to fix a minimum Q value and maximize P among decision rules with that value of Q or better.

2.3.6 Equivalence of Objective Functions

Interestingly, if P is fixed then the above four objective functions become equivalent. That is,

$$\min_D (E + P - 2PQ) = \max_D (PQ) = \max_D Q$$
$$= \min_D (\text{estimated variance of estimates}).$$

The equivalence does not hold if P is merely constrained to be above a certain value.

2.3.7 Minimize Costs of Misclassification

Let c_{21} be the cost of classifying cases from Π_2 into Π_1 and c_{12} from Π_1 into Π_2. Then, the expected cost of misclassification is

$$E(C) = c_{21} \cdot P(D_1 \cap \Pi_2) + c_{12} \cdot P(D_2 \cap \Pi_1)$$
$$= c_{21} P.(1 - Q) + c_{12} (E - PQ)$$
$$= Pc_{21} + Ec_{12} - PQ (c_{12} + c_{21})$$

For the case where P = E and Q = 1, that is, the perfect decision rule, we have:

$$E(C) = c_{21}E + c_{12}E - E (c_{12} + c_{21}) = 0$$

For the case where PQ = E,

$$E(C) = Pc_{21} - Ec_{21} = (P - E) c_{21}.$$

This is intuitively reasonable since, if PQ = E, no Π_1 cases are classified into Π_2, so the cost of misclassification depends only on the Π_2 cases classified into Π_1.

Upper bounds on the optimal costs of misclassification may be derived as follows: two trivial rules are

$D^{(1)}$: classify all cases into Π_1

$D^{(2)}$: classify all cases into Π_2

Under $D^{(1)}$, $E_{D^{(1)}}(C) = c_{21} \cdot P(D_1 \cap \Pi_2) + c_{12} \cdot P(D_2 \cap \Pi_1)$

$= c_{21}P \cdot (1-Q) = c_{21}(1-Q) = c_{21}(1-E)$ since $P = 1$ and $Q = E$.

Under $D^{(2)}$, $E_{D^{(2)}}(C) = c_{12}(E - PQ) = c_{12}E$, since $P = 0$.

$$E_{D^*}(C) = \inf_{D \in D} E_D(C) \leq \min (c_{21}(1-E), C_{12}E).$$

In the most general context, there may be costs and benefits associated with correct decisions as well as incorrect decisions. Thus, we assume there are weights w_{ij}, $i,j = 1,2$ associated with each entry in the 2 x 2 table. The expected "cost" of classification is, therefore,

$$E(w) = \sum_{i,j} w_{ij} P_{ij}, \quad i, j = 1,2, \text{ where the } P_{ij} \text{ are as defined in 2.3.}$$

In terms of the parameters P, Q and E, we have

$$E(w) = w_{11}PQ + w_{12}(E-PQ) + w_{21}(P(1-Q)) + w_{22}(1-P-(E-PQ))$$

$$= PQ(w_{11}-w_{12}-w_{21}+w_{22}) + Ew_{12}+Pw_{21} + (1-P)w_{22}-Ew_{22}.$$

For the "perfect" screen with P = E and Q = 1, we get

$$E(w) = E(w_{11}-w_{12}-w_{21}+w_{22}) + Ew_{12}+Ew_{21}+ (1-E)w_{22}-Ew_{22}$$

$$= E(w_{11}-w_{11}) + w_{22}.$$

Thus, the "perfect" screen does not yield zero cost unless $w_{11} = w_{22} = 0$. The decision rule yields a negative cost if $\frac{w_{22}}{w_{22}-w_{11}} > E$ or if $w_{22} = w_{11} < 0$.

2.3.8 Screening as a Hypothesis Testing (Decision) Problem

For each new observation $\vec{x}$ we can view the screening problem as a decision between two hypotheses:

H_0: $\vec{x} \in \Pi_1$ $\qquad$ H_A: $\vec{x} \in \Pi_2$

Thus, for a decision rule $D = \langle D_1, D_2 \rangle$, D_2 is the rejection region for H_0, with

$$\alpha(D) = P(\vec{x} \in D_2 | \vec{x} \in \Pi_1) = \frac{E-PQ}{E} = 1 - \frac{PQ}{E} \qquad (1)$$

$$\beta(D) = P(\vec{x} \in D_1 | \vec{x} \in \Pi_2) = \frac{P(1-Q)}{1-E} \qquad (2)$$

$$\alpha(D) + \beta(D) = \frac{E-PQ}{E} + \frac{P(1-Q)}{1-E} = \frac{(1-E)\ (E-PQ) + PE(1-Q)}{E(1-E)}$$

$$= \frac{E-PQ - E^2 + PQE + PE - PQE}{E\ (1-E)}$$

$$= \frac{E(1-E) - P(Q-E)}{E(1-E)} = 1 - \frac{P(Q-E)}{E(1-E)} \qquad (3)$$

If $P = E$ and $Q = 1$, $\alpha(D) + \beta(D) = 0$.

In classical hypothesis testing, the aim is to minimize $\beta(D)$ for fixed $\alpha(D) \leq \alpha^*$ where α^* is a pre-specified level. Thus, we must minimize (3) subject to (1) less than α^*, i.e.,

$$1 - \frac{PQ}{E} \leq \alpha^* \Rightarrow \frac{PQ}{E} \geq 1 - \alpha^*$$

or $PQ \geq E\ (1 - \alpha^*)$.

But $\beta(D) = 1 - \frac{P \cdot (Q-E)}{E\ (1-E)}$,

which is minimized by maximizing $P \cdot (Q-E)$.

Thus, in terms of classical hypothesis testing the problem is:

maximize $P(Q-E)$ subject to $PQ \geq E(1-\alpha^*)$ (4)

If instead, we wish to minimize the sum of $\alpha(D)$ and $\beta(D)$, then by (3) the problem is to

$$\min \left[1 - \frac{P(Q-E)}{E(1-E)}\right] \equiv \max P(Q-E) \qquad (5)$$

Note again that, for fixed P, the objective function is to maximize Q.

2.4 Screening and Prediction Logic

2.4.1 Background

Hildebrand, Laing and Rosenthal (1977) consider the use of a variable X taking on c values $x_1, \ldots, x_c$ to predict

the state of another variable Y taking on r values Y_1, ..., Y_r. They define a degree 1 proposition as being one that makes a prediction on Y for each observation on X.

If we consider the decision rule $D = \langle D_1, D_2 \rangle$ as a two-state prediction variable and $\Pi = \langle \Pi_1, \Pi_2 \rangle$ as the two-state dependent variable, then we have a simple prediction logic statement:

If $\vec{x} \epsilon D_1$, predict $\vec{x} \epsilon \Pi_1$, i.e., $D_1 \rightarrow \Pi_1$.

Similarly, $D_2 \rightarrow \Pi_2$.

That is, in Hildebrand et al's notation, we define

$$X = \begin{cases} X_1 \text{ if } D_1 \text{ occurs} \\ X_2 \text{ if } D_2 \text{ occurs} \end{cases}$$

$$Y = \begin{cases} Y_1 \text{ if } \Pi_1 \text{ is the true state} \\ Y_2 \text{ if } \Pi_2 \text{ is the true state} \end{cases}$$

with the corresponding 2 x 2 table:

	X_1	X_2	
Y_1	P_{11}	P_{12}	$P_{1.}$
Y_2	P_{21}	P_{22}	$P_{2.}$
	$P_{.1}$	$P_{.2}$	1

and the prediction logic statement $X_1 \rightarrow Y_1$, $X_2 \rightarrow Y_2$.

In prediction logic problems, rules linking X states to Y states are formulated a priori. In the screening problem, we use the data to identify decision rules (prediction variables), hence the prediction is ex post. Nonetheless, the theory of prediction logic can be used to develop alternative objective functions and methods of evaluating results. This is the subject of the remainder of this chapter.

2.4.2 Proportionate Reduction in Error

In section 2.3, we discussed some of the possible objective functions for the screening problem. In this section,

we introduce an alternative objective function that has properties that make it preferable to the others in some circumstances. We also compare this objective function with some of the well-known measures of association in 2 x 2 tables.

Costner (1965) argues that a measure of association in a contingency table should have what he calls a proportionate reduction in error interpretation using the following four criteria:

1. The user must define a rule for predicting the dependent variable given the independent variable (this is called rule K).

2. The user must define a corresponding rule for estimating the dependent variable without the independent variable (called rule U).

3. The user must define "errors" in prediction.

4. The measure must be defined in proportionate reduction in error form:

$$\frac{\text{Errors Rule U - Errors Rule K}}{\text{Errors Rule U}}$$

Thus, a proportionate reduction in error (PRE) measure reflects the proportionate reduction in error achieved by using the information in the independent variable rather than just the information in the dependent variable.

Applying the PRE concept to the screening problem we have:

Rule K: $X_1 \rightarrow Y_1$ and $X_2 - Y_2$. Defining an "error" as a misclassified case, the proportion of errors by rule K is simply the total probability of misclassification, $E + P - 2PQ$.

Rule U: In order to correspond to rule K, the rule U must classify the same proportion into Y_1 and Y_2 as rule K did. Thus, rule U randomly classifies a proportion P of the cases into Y_1 and 1-P into Y_2. The expected proportion of errors by rule U, then, is $P(1-E) + (1-P)E$, or $E + P - 2PE$. Note that this is the probability of misclassification when $Q = E$, i.e., $Q = P(Y_1|X_1) = P(Y_1) = E$, so that X_1 provides no information.

Thus, the PRE measure, denoted by ∇, is

$$\nabla = \frac{(E + P - 2PE) - (E + P - 2PQ)}{E + P - 2PE}$$

$$= \frac{2P\cdot(Q - E)}{E + P - 2PE} \qquad (1)$$

The corresponding ∇ value for the sample is

$$\hat{\nabla} = \frac{2p\cdot(q - e)}{e + p - 2pe} \qquad (2)$$

Again, maximizing ∇, for fixed P, is equivalent to maximizing Q. That is, maximizing ∇ becomes equivalent to the objective functions discussed in section 2.3. I now develop some basic results for this PRE measure. In all these results I assume, without loss of generality, that $0<E<\frac{1}{2}$.

Theorem 2.4.1: $\nabla<0 \iff Q<E$. Thus, any decision rule with $Q<E$ is inadmissible since rule U outperforms it (or the rule $X_1\to Y_2$, $X_2\to Y_1$). Furthermore, $\nabla = 0 \iff Q = E$.

Proof: $\nabla = \frac{2P\cdot(Q - E)}{E + P - 2PE} = c\,(Q - E)$

where $c = \frac{2P}{E + P\cdot(1 - 2E)} \geq 0$ since $E<\frac{1}{2}$.

Thus $\nabla<0 \iff Q<E$ as required. The second statement holds trivially.

Theorem 2.4.2: $\nabla\leq 1$ with equality $\iff$ P = E and Q = 1.

Proof: $\nabla = 1 - \frac{E + P - 2PQ}{E + P - 2PE} > 1$

$$\Rightarrow \frac{E + P - 2PQ}{E + P - 2PE} < 0$$

$\Rightarrow$ E + P - 2PQ = Probability of misclassification <0

which is impossible. Hence, $\nabla\leq 1$

$\nabla = 1 \Rightarrow E + P - 2PQ = 0$

$\Rightarrow (E - PQ) + (P - PQ) = 0$

$\Rightarrow$ E = PQ and P = PQ (since $E\geq PQ$ and $P\geq PQ$)

$\Rightarrow$ E = P = PQ

$\Rightarrow$ Q = 1

The converse holds also. If Q = 1 and P = E,

$$\nabla = \frac{2E\ (1 - E)}{E + E - 2E^2} = 1.$$

Thus, Theorems 2.4.1 and 2.4.2 show that, for any admissible decision rules, $0 \leq \nabla \leq 1$, with $\nabla = 0$ when Q = E (the decision rule provides no additional information) and $\nabla = 1$ when Q = 1 and P = E (the perfect decision rule).

Theorem 2.4.3: If $D^{(1)}$ and $D^{(2)}$ are two admissible decision rules with (P_1, Q) and (P_2, Q) the respective parameters, then $\nabla_{D^{(1)}} > \nabla_{D^{(2)}} \iff P_1 > P_2$.

Proof: $\nabla_{D^{(1)}} > \nabla_{D^{(2)}} \Rightarrow \frac{2P_1 \cdot (Q - E)}{E + P_1(1-2E)} > \frac{2P_2 \cdot (Q - E)}{E + P_2(1-2E)}$

$\Rightarrow P_1 \left[E + P_2\ (1 - 2E\right] > P_2 \left[E + P_1\ (1 - 2E)\right]$
since $Q > E$ and $0 < E < \frac{1}{2}$

$\Rightarrow P_1E + P_1P_2 - 2P_1P_2E > P_2E + P_1P_2 - 2P_1P_2E$

$\Rightarrow P_1 > P_2$ as required.

Reversing the above steps, the converse holds.

Theorem 2.4.4: If $D = \langle D_1, D_2 \rangle$ is a decision rule and ∇_{D_1} and ∇_{D_2} are the PRE measures for the components of that decision rule, then,

$$\nabla_D = \frac{(\text{Rule } U_{D_1})\ \nabla_{D_1} + (\text{Rule } U_{D_2})\ \nabla_{D_2}}{\text{Rule } U_{D_1} + \text{Rule } U_{D_2}}$$

That is, the PRE measure for the decision rule D can be derived as the weighted average of the PRE measure for each component.

Proof: D_1 is the equivalent to the prediction logic statement $X_1 \rightarrow Y_1$.

Rule K_{D_1}: Under D_1, errors occur only for cases (Y_2, X_1) which occur with frequency $P \cdot (1-Q)$.

Rule U_{D_1}: Under D_1, predictions are made for a proportion P of the cases with errors occurring with probability 1-E so that rule U errors are $P \cdot (1-E)$

$$\nabla_{D_1} = \frac{P \cdot (1 - E) - P \cdot (1-Q)}{P\ (1 - E)} = \frac{Q - E}{1 - E}$$

D_2 is equivalent to the statement $X_2 \rightarrow Y_2$.

Rule K_{D_2}: Under D_2, predictions are made for the cases (Y_1, X_2) which occur with frequency E - PQ.

Rule U_{D_2}: Under D_2, Rule U predictions are made for a proportion 1 - P of the cases with errors occurring with probability E, so Rule U errors are (1 - P)E.

$$\therefore \nabla_{D_2} = \frac{(1 - P)E - (E - PQ)}{(1 - P)E} = \frac{P\,(Q - E)}{(1 - P)E}$$

$$\text{Now } \frac{(\text{Rule } U_{D_1})\ \nabla_{D_1} + (\text{Rule } U_{D_2})\ \nabla_{D_2}}{\text{Rule } U_{D_1} + \text{Rule } U_{D_2}} =$$

$$= \frac{P\cdot(1-E)\left(\frac{Q-E}{1-E}\right) + (1-P)E\,\frac{P\cdot(Q-E)}{(1-P)E}}{P\cdot(1-E) + (1-P)E} = \frac{2P\cdot(Q-E)}{E+P-2PE}$$

$$= \nabla_D \text{ as required.}$$

2.4.3 Comparison to Other Measures

2.4.3.1 Guttman's λ

Consider the following PRE measure:

Rule U: Predict the most likely value. Since we assume $E<\frac{1}{2}$, this means we predict $Y_2(\Pi_2)$ for every case and make errors with probability E.

Rule K: Predict the most likely Y value given the X value. For our situation, this means $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$, with errors = probability of misclassification = E + P - 2PQ.

$$\text{Then } \nabla_\lambda = \frac{E - (E + P - 2PQ)}{E} = \frac{2PQ - P}{E} \qquad (1)$$

Guttman's λ is given by

$$\lambda = \frac{\Sigma M_j - M}{1 - M} \qquad (2)$$

where $M_j = \max_i P_{ij}$

$M = \max_i P_i$.

In our notation,

$$M_1 = P_{11} = PQ$$

$$M_2 = P_{22} = (1 - P) - (E - PQ)$$

$$M = 1 - E$$

$$\therefore \lambda = \frac{PQ + (1 - P) - (E - PQ) - (1 - E)}{1 - (1 - E)}$$

$$= \frac{PQ + 1 - P + PQ - 1}{E} = \frac{2PQ - P}{E} = \nabla_\lambda \qquad (2)$$

Thus, we have demonstrated that Guttman's λ has a PRE interpretation.

Theorem 2.4.5: $\nabla = 1 \iff \nabla_\lambda = \lambda = 1$
$\iff Q = 1$ and $P = E$

Proof: We have already shown (Theorem 2.4.2) that $\nabla = 1 \iff Q = 1$ and $P = E$. Therefore we need only show $\lambda = 1 \iff Q = 1$ and $P = E$.

If $Q = 1$ and $P = E$, $\lambda = \frac{2E - E}{E} = 1$.

If $\lambda = 1$, then using (2) we have $\frac{P_{11} + P_{22} - P_{2\cdot}}{1 - P_{2\cdot}} = 1$

$\Rightarrow P_{11} + P_{22} = 1$
$\Rightarrow P_{21} + P_{12} = 0$
$\Rightarrow P_{21} + P_{12} = 0$
$\Rightarrow P\cdot(1 - Q) = 0$ and $E - PQ = 0$
$\Rightarrow P = E$ and $Q = 1$ as required.

Thus, λ shares the property that it equals one only if the decision rule is perfect.

Also,

$\lambda = 0 \iff q = \frac{1}{2}$.

$\lambda < 0 \iff q < \frac{1}{2}$.

This is consistent with the PRE derivation of λ since, when $q = \frac{1}{2}$, there is no modal class for X_1. When $q < \frac{1}{2}$, the modal class has not been selected given X_1 which means that the prediction should have been $X_1 \rightarrow Y_2$, for which $\lambda \geq 0$. Thus, if the

decision rule is selected according to Rule K,

$$0 \leq \lambda \leq 1.$$

Note also that, for fixed P, maximizing λ is equivalent to maximizing Q.

2.4.3.2 Goodman and Kruskal's τ

Consider the following PRE development:

Rule U: Predict Y_1 for a proportion E of the cases and Y_2 for a proportion 1 - E of the cases. The error rate will be $E(1-E) + (1-E)E = 2E(1-E)$.

Rule K: Predict Y_1 for a proportion $P_{11}/P_{.1}$ of the cases and Y_2 for a proportion $P_{21}/P_{.1}$ when X_1 occurs; predict Y_2 for a proportion $P_{22}/P_{.2}$ of the cases and Y_1 for a proportion $P_{12}/P_{.2}$ when X_2 occurs.

Expected error rate under Rule K is:

$$P_{.1}\left(\frac{P_{11}}{P_{.1}}\frac{P_{21}}{P_{.1}} + \frac{P_{21}}{P_{.1}}\frac{P_{11}}{P_{.1}}\right) + P_{.2}\left(\frac{P_{12}}{P_{.2}}\frac{P_{22}}{P_{.2}} + \frac{P_{22}}{P_{.2}}\frac{P_{12}}{P_{.2}}\right)$$

$$= 2\left(\frac{P_{11}\,P_{21}}{P_{.1}} + \frac{P_{22}\,P_{12}}{P_{.2}}\right)$$

$$= 2\left(\frac{PQ \cdot P(1-Q)}{P} + \frac{(E-PQ)(1-P-(E-PQ))}{1-P}\right)$$

$$= PQ(1-Q) + (E-PQ) - \frac{(E-PQ)^2}{1-P}$$

$$= \frac{(-PQ^2 + E)(1-P) - E^2 + 2EPQ - P^2Q^2}{1-P}$$

$$= \frac{-PQ^2 + E - PE - E^2 + 2EPQ}{1-P}$$

$$= \frac{E(1-E) - P(Q^2 - 2EQ + E)}{1-P}$$

$$\nabla_\tau = 1 - \frac{\dfrac{E(1-E) - P(Q^2 - 2EQ + E)}{1-P}}{E(1-E)}$$

$$= \frac{(1-P)E(1-E) - E(1-E) + P(Q^2 - 2EQ + E)}{(1-P)E(1-E)}$$

$$= \frac{P\,(Q^2 - 2EQ + E) - P\,(E\,(1 - E))}{(1 - P)\,E\,(1 - E)}$$

$$= \frac{P\,(Q^2 - 2EQ + E^2)}{(1 - P)\,E\,(1 - E)} = \frac{P\,(Q - E)^2}{(1-P)E(1-E)} \qquad (1)$$

Goodman and Kruskal's τ is defined as

$$\tau = \frac{\sum_i \sum_j \left(\frac{P_{ij}^2}{P_{.j}}\right) - \sum_i P_{i.}^2}{1 - \sum_i P_{i.}^2} \qquad (2)$$

Now, $1 - \sum_i P_{i.}^2 = 1 - E^2 - (1 - E)^2 = 2E\,(1 - E)$ (3)

The numerator of (2) is

$$\frac{P_{11}^2}{P_{.1}} + \frac{P_{12}^2}{P_{.2}} + \frac{P_{21}^2}{P_{.1}} + \frac{P_{22}^2}{P_{.2}} - P_{1.}^2 - P_{2.}^2 =$$

$$\frac{(PQ)^2}{P} + \frac{(E-PQ)^2}{1-P} + \frac{[P(1-Q)]^2}{P} + \frac{[(1-P)-(E-PQ)]^2}{1-P} - E^2-(1-E)^2$$

$$= PQ^2 + P(1-Q)^2 + \frac{2(E-PQ)^2}{1-P} + (1-P)-2(E-PQ)-E^2-(1-E)^2$$

$$= 2PQ^2 + 1 + 2\,\frac{(E-PQ)^2}{1-P} - 2E - E^2 - (1-E)^2$$

$$= 2PQ^2 + 2\,\frac{(E-PQ)^2}{1-P} - 2E + 2E\,(1-E)$$

$$= \frac{2PQ^2\,(1-P) + 2\,(E-PQ)^2 - 2E^2\,(1-P)}{(1 - P)}$$

$$= \frac{2[PQ^2 - P^2Q^2 + E^2 - 2EPQ + P^2Q^2 - E^2 + PE^2]}{1 - P}$$

$$= \frac{2\,PQ^2 - 2EPQ + PE^2}{1 - P} = \frac{2P\,(Q^2 - 2EQ + E^2)}{1 - P} \qquad (4)$$

Dividing (4) by (3), we have

$$\tau = \frac{2P\,(Q^2 - 2EQ + E^2)}{(1-P)\,E\,(1-E)} = \frac{P\,(Q - E)^2}{(1-P)E(1-E)} \qquad (5)$$

But this is the same as (1). That is,

$$\tau = \nabla_{\tau} = \frac{P \cdot (Q-E)^2}{(1-P) \cdot E \cdot (1-E)} \qquad (6)$$

Therefore, τ has a PRE interpretation.

Note that, if P = E and Q = 1,

$$\tau = \frac{E\ (1-E)^2}{(1-E)\ E\ (1-E)} = 1$$

If Q = E, τ = 0.

Also, $\tau \geq 0$ since all terms in (7) are greater than zero.

2.4.3.3 Extension to Weighted Errors

Let $W_{12} > 0$ and $W_{21} > 0$ be the weights (costs) associated with prediction errors. Then,

Rule K Expected Cost of Errors: $W_{12}P_{12} + W_{21}P_{21}$

$= W_{12}\ (E - PQ) + W_{21}P\ (1 - Q)$

Rule U Expected Cost of Errors: $W_{12}P_{\cdot 1}\ P_{2\cdot} + W_{21}P_{\cdot 2}\ P_{1\cdot}$

$= W_{12}E\ (1-P) + W_{21}\ (1-E)\ P$

Then $\nabla_{D,W} = \frac{\text{Rule U - Rule K}}{\text{Rule U}} =$

$$\frac{W_{12}E - W_{12}EP + W_{21}P - W_{21}EP - W_{21}E + W_{12}PQ - W_{21}P + W_{21}PQ}{W_{12}E - W_{12}PE + W_{21}P - W_{21}EP}$$

$$= \frac{PQ\ (W_{12} + W_{21}) - PE\ (W_{12} + W_{21})}{EW_{12} + PW_{21} - PE\ (W_{12} + W_{21})}$$

$$= \frac{P\ (Q - E)\ (W_{12} + W_{21})}{EW_{12} + PW_{21} - PE\ (W_{12} + W_{21})} \qquad (1)$$

Properties of $\nabla_{D,W}$ include:

1. For $W_{12} = W_{21} = W$, say, $\nabla_{D,W} = \nabla_D$

   Proof: $\nabla_{D,W} = \frac{P\ (Q - E)\ 2W}{EW + PW - 2PEW} = \frac{2P\ (Q - E)}{E + P - 2PE} = \nabla_D$

2. $\nabla_{D,W}$ is invariant under constant multiplication of weights: that is, $\nabla_{D,kW} = \nabla_{D,W}$

Proof: $\nabla_{D,kW} = \dfrac{P\,(Q-E)\,(kW_{12} + kW_{21})}{EkW_{12} + PkW_{21} - PE\,(kW_{12} + kU_{21})} = \nabla_{D,W}$

(However $\nabla_{D,W}$ is not invariant under addition.)

3. $\nabla_{D,W} = 0 \iff Q = E$

Proof: Obvious from (1).

4. $\nabla_{D,W} = 1 \iff Q = 1$ and $P = E$

Proof: If $Q = 1$ and $P = E$, $\nabla_{D,W} =$

$$\frac{E\,(1 - E)\,(W_{12} + W_{21})}{EW_{12} + EW_{21} - E^2\,(W_{12} + W_{21})}$$

$$= \frac{E\,(1 - E)\,(W_{12} + W_{21})}{(E - E^2)\,(W_{12} + W_{21})} = 1.$$

If $\nabla_{D,W} = 1$, then $\dfrac{\text{Rule U-Rule K}}{\text{Rule U}} = 1 \Rightarrow \dfrac{\text{Rule K}}{\text{Rule U}} = 0$

$\Rightarrow$ Rule K $= 0 \Rightarrow W_{12}(E-PQ) + W_{21}P(1-Q) = 0$

$\Rightarrow E-PQ = 0$ and $P(1-Q) = 0$ (since $W_{12}, W_{21} > 0$)

$\Rightarrow P = E$ and $Q = 1$.

Thus, $\nabla_{D,W}$ is a straightforward extension of ∇_D and, as such, is a preferred objective function for any problem where there is a difference in costs associated with the two decisions. Furthermore, by property 2., the user need only specify the relative magnitude of the costs.

2.5 Statistical Inference

In the preceding sections, we have introduced objective functions that may be relevant to the screening problem. However, in practice we must deal with sample estimates of the objective function. Thus, questions of statistical inference arise.

In the material that follows, we consider the PRE measure

$$\nabla_D = \frac{EP\,(Q - E)}{E + P - 2PE}$$

and its sample based estimator

$$\hat{\nabla}_D = \frac{2p\ (q - e)}{e + p - 2pe}$$

The results are easily extended to $\nabla_{D,W}$ and $\hat{\nabla}_{D,W}$.

Recall that we are assuming a random sample from the population $\Pi = (\Pi_1, \Pi_2)$ so that the true marginal probabilities are unknown for both variables X and Y and neither set of marginal totals is fixed (the sample is not stratified). The sample results may be written as follows:

	D_1	D_2	
Π_1	p_{11}	p_{12}	$p_{1\cdot}$
Π_2	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

2.5.1 Estimated Variance of $\hat{\nabla}_D$

In terms of the above table,

$$\hat{\nabla} = 1 - \frac{p_{12} + p_{21}}{p_{1\cdot}\, p_{\cdot 2} + p_{2\cdot}\, p_{\cdot 1}}$$

For sufficiently large n, each of the p_{ij}, $p_{i\cdot}$, $p_{\cdot j}$ follow an approximately normal distribution with means P_{ij}, $P_{i\cdot}$, $P_{\cdot j}$ respectively and variance

$$\sqrt{\frac{P_{ij}\ (1-P_{ij})}{n}} \quad , \quad \sqrt{\frac{P_{i\cdot}\ (1-P_{i\cdot})}{n}} \quad , \quad \sqrt{\frac{P_{\cdot j}\ (1-P_{\cdot j})}{n}} \ .$$

Thus, $\hat{\nabla}$ is a well-behaved function of variables that each asymptotically follow a normal distribution with variance approaching zero. Thus $\hat{\nabla}$ itself follows an asymptotic normal distribution with variance estimated by using the Taylor expansion for ∇, i.e.,

$$\nabla = c + \sum_{ij}\sum - a_{ij}\, P_{ij} \qquad (1)$$

where

$$- a_{ij} = \frac{\partial \nabla}{\partial P_{ij}} .$$

$$\frac{\partial \nabla}{\partial P_{11}} = -\frac{\partial\left(\frac{P_{12} + P_{21}}{P_1.P._2 + P_2.P._1}\right)}{\partial P_{11}} = \frac{(P_{12}+P_{21})\ \frac{\partial}{\partial P_{11}}\ (P_1.P._2 + P_2.P._1)}{(P_1.P._2 + P_2.P._1)^2}$$

$$= \frac{(P_{12} + P_{21})\ (P._2 + P_2.)}{(P_1.P._2 + P_2.P._1)^2} = (1 - \nabla)\ \frac{(P._2 + P_2.)}{(P_1.P._2 + P_2.P._1)}$$

$$= \frac{(1 - \nabla)}{U}\ (1 - P + 1 - E) = \frac{1 - \nabla}{U}\ (2 - P - E) = -a_{11} \quad (2)$$

where $U = P_1.P._2 + P_2.P._1$ = Rule U error rate.

Similarly,

$$\frac{\partial \nabla}{\partial P_{12}} = -\frac{1}{U} + \frac{P_{12} + P_{21}}{U^2}\ (P_2. + P._1) = -\frac{1}{U} + \frac{1 - \nabla}{U}\ (1-P+E)$$

$$= \frac{(1 - \nabla)\ (1 - P + E) - 1}{U} \quad (3)$$

$$\frac{\partial \nabla}{\partial P_{21}} = -\frac{1}{U} + \frac{P_{12} + P_{21}}{U^2}\ (P_2. + P._1)$$

$$= -\frac{1}{U} + \frac{1 - \nabla}{U}\ (P + 1 - E)$$

$$= -\frac{1}{U} - \left(\frac{1 - \nabla}{U}\right)\ (1 - P + E) \quad (4)$$

$$\frac{\partial \nabla}{\partial P_{22}} = \frac{(P_{12} + P_{21})}{U^2}\ (P_1. + P._1) = \frac{(1 - \nabla)}{U}\ (P + E) \quad (5)$$

Then $\text{var}\ (\hat{\nabla}) \cong \frac{1}{n}\ \left[\sum\sum_{ij} a_{ij}^2\ P_{ij} - \left(\sum\sum_{ij} a_{ij}\ P_{ij}\right)^2\right]$ (6)

with $a_{ij} = -\frac{\partial \nabla}{\partial P_{ij}}$.

Substituting the sample estimates, we have

$$\hat{\text{var}}\ (\hat{\nabla}) = \frac{1}{n}\ \left(\sum\sum_{ij} \hat{a}_{ij}^2\ P_{ij} - \left(\sum\sum_{ij} a_{ij}\ P_{ij}\right)^2\right) \quad (7)$$

$$\hat{\text{var}}\ (\hat{\nabla}) = \hat{A}'\hat{\Sigma}\ \hat{A} \quad (8)$$

where $A = \begin{pmatrix} \hat{a}_{11} \\ \hat{a}_{12} \\ \hat{a}_{21} \\ \hat{a}_{22} \end{pmatrix}$, $\hat{\Sigma} = (\text{cov}\ p_{ij}, p_{i'j})$

2.5.2 Hypothesis Tests and Confidence Intervals

Given $\hat{\nabla}$ and $\hat{var}\ (\hat{\nabla})$, the asymptotic normality of ∇ allows confidence intervals to be formed in the usual way, i.e., the 100 $(1 - \alpha)$% confidence interval is given by

$$\hat{\nabla} \pm Z_{\alpha/2} \sqrt{var\ (\hat{\nabla})}. \qquad (1)$$

where $Z_{\alpha/2}$ is the standard normal deviate for probability $\alpha/2$.

To test the hypothesis H_0: $\nabla = 0$ against the one-sided alternative $\nabla > 0$, we reject H_0 if $\hat{\nabla} > Z_\alpha \sqrt{var\ (\nabla)}$ where Z_α is the standard normal deviate for type one error = α.

To compare two independent $\hat{\nabla}$ values (e.g., $\hat{\nabla}$ for two independent samples, or for a before and after comparison), use

$$Z = \frac{\hat{\nabla}_1 - \hat{\nabla}_2}{var\ (\hat{\nabla}_1) + var\ (\hat{\nabla}_2)} \qquad (2)$$

and reject H_0: $\nabla_1 - \nabla_2 = 0$ if $|Z| \geq Z_{\alpha/2}$.

In the before vs. after comparison, a one-sided test might be appropriate, i.e.,

$$H_0: \ \nabla_1 \leq \nabla_2 \text{ vs. } H_a: \nabla_1 > \nabla_2.$$

Reject H_0 if $Z > Z_\alpha$.

Note: As usual, the normal approximation may be improved by replacing $\hat{\nabla}$ by $\hat{\nabla}_+ = 1 - \left(\frac{\hat{k} + 1/n}{\hat{u}}\right)$ or $\hat{\nabla}_- = 1 - \left(\frac{\hat{k} - 1/n}{\hat{u}}\right)$

For example the confidence interval becomes

$$\hat{\nabla}^+ - Z\alpha_{/2} \sqrt{\hat{var}(\nabla)}, \ \hat{\nabla}^- + Z\alpha_{/2} \sqrt{\hat{var}(\nabla)}.$$

To test the validity of the normal approximation the usual rules apply:

$5 < nk$ and $n\ (1-k)$, where k = rule k error rate

= probability of misclassification.

Hildebrand et al (1977) conducted some Monte Carlo experiments with $\hat{\nabla}$ and $\hat{var}\ (\hat{\nabla})$ with the following general conclusions:

1) The bias of $\hat{\nabla}$ is small, particular for n>100. Bias seems to depend on the skewness of the marginal. For the case of 2 x 2 tables, they found the bias to be negative, i.e., $\hat{\nabla}$ is a conservative estimator.

2) $\hat{Var}(\hat{\nabla})$ appears to be a good approximation, although generally conservative.

3) $\hat{Var}(\hat{\nabla})$ is seriously biased, usually negatively, for small samples, especially when nk<5. However, the continuity correction helps adjust for the bias.

2.6 Statistical Inference for Ex Post Analysis

The decision rules D are based on analysis of a sample from the joint population $\Pi = \langle \Pi_1, \Pi_2 \rangle$. Thus, since the rules are not selected a priori, there is the danger that rules with a high value of $\hat{\nabla}$ are fitting the data rather than the underlying situation. In particular, high values of $\hat{\nabla}$ will tend to be optimistically biased, especially for small sample sizes.

In practice, the potential bias may be minimized by the following:

1) Choosing only rules D which make sense in the light of intuitive knowledge about the populations Π_1 and Π_2. That is, subjective knowledge could be used to help choose between rules with similar values of $\hat{\nabla}$.

2) Avoiding rules that involve too many variables. That is, for two rules $D^{(1)}$ and $D^{(2)}$ with $\hat{\nabla}_{D^{(1)}} \cong \hat{\nabla}_{D^{(2)}}$, choose the rule with the fewer variables.

More technical methods include:

1) Develop the decision rule on one portion of the sample, then test it on the remainder of the sample. The test of statistical difference described in 2.5 could be used.

2) Use a jacknife type procedure, wherein $\hat{\nabla}_{-j}$ is the result when developing the rule on all observations except the jth, with

$$\hat{\nabla}^* = n\hat{\nabla} - \frac{n-1}{n} \sum_{j=1}^{n} \hat{\nabla}_{-j}$$

(See Miller, R. G. [1974]: "The Jacknife - A Review," Biometrika, 61, 1-15.)

3) Develop the rule on one sample, then test the results on a new sample from the population. Although this is similar to 1), there are practical differences in this approach. See Chapter 4 for an application.

4) Develop hypothesis tests that take into account the Ex Post nature of the analysis.

In this section we consider approach 4). To do this, we make use of some results of Hildebrand et al. Specifically, they suggest that, for Ex Post analysis, the correct hypothesis is

$$H_O: \text{ No } \nabla_D > O \text{ vs } H_A: \text{ Some } \nabla_D > O.$$

The hypothesis is highly restrictive in the context of screening since it implies statistical independence between the dependent variables and any combination of the independent variables.

To develop the test statistic, the variance of $\hat{\nabla}$ must be computed under the assumption of statistical independence. Hildebrand et al (p. 223) derive this variance as:

$$\frac{1}{(n-1)U^2}\left(\sum_{ij}\sum W_{ij}^2 P_{i.}P_{.j} - \sum_i \pi_{i.}^2 P_{i.} - \sum_j \pi_{.j}^2 P_{.j} + U^2\right) \quad (1)$$

Where $\pi_{i.} = \sum_j W_{ij} P_{ij}$

$\pi_{.j} = \sum_i W_{ij} P_{i.}$

For the case $W_{11} = W_{22} = O$; $W_{12} = W_{21} = 1$, the expression can be simplified considerably:

$$\pi_{1.} = P_{.2} = 1 - P$$

$$\pi_{2.} = P_{.1} = P$$

$$\pi_{.1} = P_{2.} = 1 - E$$

$$\pi_{.2} = P_{1.} = E$$

Thus (1) becomes

$$\frac{1}{(n-1)U^2}\left\{P_{1.}P_{.2} + P_{2.}P_{.1} - (1-P)^2 P_{1.} - P^2 P_{2.} - (1-E)^2 P_{.1} - E^2 P_{.2} + U^2\right\}$$

$$= \frac{1}{(n-1)U^2}\left\{E(1-p)+(1-E)P - (1-P)^2 E - P^2(1-E) - (1-E)^2 P - E^2(1-P) + U^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{E-EP+P-PE^2-E+2PE-P^2E-P^2+P^2E-P+2EP-E^2P-E^2+PE^2+U^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{-P^2 + 2EP - E^2 + U^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{U^2 - (P - E)^2\right\}$$

$$= \frac{1}{n-1} \left\{1 - \left(\frac{P - E}{U}\right)^2\right\} \qquad (2)$$

Another expression for the variance can be developed as follows:

$$\text{Variance} = \frac{1}{(n-1)} \left\{1 - \left(\frac{P - E}{U}\right)^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{U^2 - (P - E)^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{(E + P - 2PE)^2 - (P - E)^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{E^2 + P^2 + 4P^2E^2 + 2EP - 4P^2E - 4PE^2 - P^2 + 2EP - E^2\right\}$$

$$= \frac{1}{(n-1)U^2} \left\{+4P^2E^2 + 4EP - 4P^2E - 4PE^2\right\}$$

$$= \frac{4}{(n-1)U^2} \left\{EP\ (PE + 1 - P - E)\right\}$$

$$= \frac{4}{(n-1)U^2} \left\{EP\ (1 - E - P\ (1 - E))\right\}$$

$$= \frac{4}{(n-1)U^2} \left\{E\ (1 - E)\ P\ (1 - P)\right\}$$

$$= \frac{4}{(n-1)U^2} \left\{\text{Product of marginals}\right\} \qquad (3)$$

Note that the variance expression depends only on P and E. This is not surprising, of course, since under the hypothesis of independence, Q = E.

The test statistic is, then,

$$\hat{Z}^2 = \frac{\hat{\nabla}^2}{\hat{\text{var}}_D \hat{\nabla}} \qquad (4)$$

where $\hat{var}_{p\,\hat{\nabla}}$ is expression (2) or (3) with p and e replacing P and E. Hildebrand et al show that $\frac{n}{n-1}\hat{Z}^2$ is approximately χ^2 with (R-1) (C-1) degrees of freedom. As discussed previously, however, the degrees of freedom for the screening problem with k variables X_i each taking on s_i values will be large and, in nearly every application, greater than 30. Therefore, we reject H_0 if

$$\hat{Z}^2 > \chi^2_\alpha (\alpha)$$

where α is the specified type 1 error.

This test is highly conservative in that it would reject the hypothesis of independence much less frequently than warranted. For instance, if we assumed the decision rule had been selected a priori, a standard χ^2 test for independence in a 2 x 2 table would have been used with only 1 degree of freedom. For the Chi-square distribution, $\chi^2_m (\alpha) > \chi^2_1 (\alpha)$ for any integer m>1.

2.7 Summary

In this Chapter, we have shown the following:

- Any decision rule for screening can be characterized by the parameters P, Q and E.
- There are a number of objective functions, each of which can be described in terms of the above parameters, that can be considered. However, a Proportionate Reduction in Error objective function has intuitive appeal.
- The screening problem can be described as an Ex Post search for a decision rule that makes good predictions in the context of prediction logic.
- Confidence intervals and hypothesis tests for the PRE measure are developed, including a conservative test for the significance of a decision rule selected Ex Post.

However, we have not yet discussed some of the techniques for developing decision rules based on sample data. In the following Chapter, we review some of the well known techniques and, in so doing, provide a rationale for the new techniques discussed in Chapters 4 and 5.

III. REVIEW OF OTHER TECHNIQUES

III. REVIEW OF OTHER TECHNIQUES

In this Chapter, we review some of the well known statistical screening methodologies. The purpose of this review is to identify some of the deficiencies associated with these techniques, especially for practical problems dealing with qualitative variables. Based on this review, we define five general properties that a desirable screening technique should possess. These properties are then used to motivate the new statistical screening techniques discussed in Chapters IV and V.

The approaches considered in this Chapter are:

- Linear Discriminant Analysis
- Regression Analysis Approach
- Logit and Probit Analysis
- Multinomial Models
- Automatic Interaction Detection

3.1 Linear Discriminant Function

Fisher's Linear Discriminant Function (LDF), developed in 1935, is both the foundation of and most prevalent of statistical screening techniques. For this reason, I will use the LDF as the basis for establishing properties that a statistical screening technique should have, particularly when dealing with qualitative variables.

In terms of decision rules $D = \langle D_1, D_2 \rangle$, the optimal partition to minimize the probability of misclassification for the case where Π_1 and Π_2 are multivariate normal $N(\vec{U}_1, \mathbf{\Sigma})$, $N(\vec{U}_2, \mathbf{\Sigma})$ with prior probabilities E and 1-E respectively is,

$$D_1^* = \left\{ \vec{x} \mid \left(\vec{U}_1 - \vec{U}_2\right)' \mathbf{\Sigma}^{-1} \left(\vec{x} - \tfrac{1}{2} \left(\vec{U}_1 + \vec{U}_2\right)\right) \geq \frac{1 - E}{E} \right\} \quad (1)$$

$$D_2^* = \left\{ \vec{x} \mid \cdot < \frac{1 - E}{E} \right\}$$

Replacing $\vec{U}_1$, $\vec{U}_2$ and $\mathbf{\Sigma}$ by the usual MLE, we obtain the classification rule:

$$\hat{D}_1^* = \left\{ \vec{x} \left(\vec{x}_1 - \vec{x}_2\right)' S^{-1} \left(\vec{x} - \tfrac{1}{2} \left(\hat{U}_1 + \hat{U}_2\right)\right) \geq \frac{1 - e}{e} \right\} \quad (2)$$

Although (1) is optimal when the assumptions of normality and equal covariance hold, optimality cannot be claimed when there are departures from the assumptions. In particular, the

assumption of equal covariance matrices tends to be restrictive. In such cases, a quadratic discriminant function arises instead of the linear function of (1). For qualitative variables, the assumption of multivariate normality will be violated except for large samples.

Goldstein and Dillon (1977) present an example that demonstrates the inappropriateness of the LDF for qualitative variables:

$$\text{Let } X_1 = \begin{cases} 0 \text{ if birth weight is low} \\ 1 \text{ if birth weight is high} \end{cases}$$

$$X_2 = \begin{cases} 0 \text{ if gestation length is short} \\ 1 \text{ if gestation length is long} \end{cases}$$

Normal babies have high birth weight and long gestation length or low birth weight and short gestation length. Abnormal babies have either of the other two combinations ((0, 1) or (1, 0)).

The LDF decision rule is of the form:

if $B_1X_1 + B_2X_2 \geq c$, classify in Π_1 (normal group)

if $B_1X_2 + B_2X_2 < c$, classify in Π_2 (abnormal group).

For the rule to classify correctly, we must have

$B_1X_1 + B_2X_2 \geq c$ for (0, 0) and (1, 1) and

$B_1X_1 + B_2X_2 < c$ for (0, 1) and (1, 0)

Thus:

$0 \geq c$ for (0, 0)

$B_1 + B_2 \geq c$ for (1, 1)

but $B_2 < c$ for (0, 1)

and $B_1 < c$ for (1, 0)

$\therefore B_1 + B_2 < 2c < c < 0$, unless $c = 0 \Rightarrow B_1 = B_2 = 0$.

This result can be generalized by considering the likelihood function

$$L(\vec{x}) = B_0 + \sum_{j=1}^{P} c_j x_j \qquad (1)$$

Now consider any pair (X_1, X_2), say. From (1),

$$L(1,1,X_3,\ldots,X_p)=L(0,1,X_3,\ldots,X_p) + L(1,0,X_3,\ldots,X_p) - L(0,0,X_3,\ldots,X_p)$$

or $$L(1,1) = L(0,1) + L(1,0) - L(0,0) \qquad (2)$$

If $L(0,0) < \min(L(0,1), L(1,0))$,

then $$L(1,1) > L(0,1) + L(1,0) - \min(L(0,1), L(1,0))$$

$$= \max(L(0,1), L(1,0)).$$

Thus, if $(0,0)\varepsilon\Pi_1$ and $(0,1), (1,0)\varepsilon\Pi_2$, then $(1,1)$ could not be classified into Π_1.

For qualitative variables where the categories have no particular order, linear combinations of variables may be meaningless. Although the LDF need not perform poorly, there is always the danger it could do so in examples of the above type. Clearly, a random reordering of the categories of the variables could significantly effect the LDF and its success in classification. For example, by redefining X_1 so that

$$X_1 = \begin{cases} 1 & \text{if birth weight is low} \\ 0 & \text{if birth weight is high} \end{cases}$$

while keeping the same definition for X_2, it is no longer impossible to classify all possibilities correctly.

Based on this observation, I suggest that a desirable property of a statistical method for screening with qualitative variables is:

Property 1: The statistical method should not be affected by random reordering of categories of each variable.

The LDF is also restricted in terms of the objective functions it can consider. That is, it is set up to minimize the total probability (or costs) of misclassification. The LDF cannot be used to minimize ∇ which, as we have seen, is a better criterion when predictability is the major concern. Also, the LDF is not amenable to maximizing the probability of misclassification subject to fixed resources, i.e., $P \leq P^*$, although users often apply it in this way. We contend, however, that this procedure involves a potentially erroneous assumption, as discussed below.

In many practical applications, the LDF is developed from the sample without any constraints. Then, the function is applied to each observation to yield a "score". The observations may be ranked from highest to lowest based on this score, i.e., $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ with scores c_1 to c_n. If P* is the desired P value, a cutoff score is found such that there at most nP* values above the score and at least $n\cdot(1-P^*)$ values below it. For example if n = 100 and P* = 0.2, c* should be chosen so that $c_{20} > c^* > c_{21}$, e.g.,

$$c^* = \frac{c_{20} + c_{21}}{2}.$$

Although this appears to be a reasonable procedure, it may not be optimal. This is because it involves the implicit assumption that there is a monotonic relationship between the score and the probability of membership in Π_1, i.e., for c*>c

$$P\left(\Pi_1 | c\ (\vec{x}) > c^*\right) \geq P\left(\Pi_1 | c\ (\vec{x}) > c\right),$$

where c $(\vec{x})$ denotes the score of $\vec{x}$ on the LDF. Put another way, the assumption is, for c*>c,

$$\text{if } D_1^* = \left\{\vec{x} | c(\vec{x}) > c^*\right\} \text{ and}$$

$$D_1 = \left\{\vec{x} | c(\vec{x}) > c\right\},$$

then $Q_{D_1^*} \geq Q_{D_1}$, where $Q_D = P(\Pi_1 | \vec{x} \varepsilon D)$. The assumption may be criticized on two counts:

1) For qualitative variables, linear combinations of variable values have no real meaning, so that a higher score on the LDF need not imply greater likelihood of membership in Π_1 (the example of birth weight is one such case).

2) Even if the assumption holds, the region D_1^* may not be optimal among regions for which $P(D_1) \leq P^*$.

Note that the P* value could also be achieved via a second procedure by randomly allocating P*/P (assuming P*<P) of the observations $\vec{x}$ for which $c(\vec{x})>c$ into Π_1. This procedure would yield the same value of Q. As we have seen, for fixed Q, $\nabla_D > \nabla_{D^*}$ if P>P*. More generally, there is a tradeoff between P and Q so that if P decreases, Q should increase and vice versa. A procedure that reduces P while fixing Q cannot be considered appropriate.

In sum, the LDF approach does not lend itself to "control" of the decision rule parameters P and Q. In practical problems,

this control may be quite important. That is, the user should be able to specify the objective function appropriate to the problem before selecting a screening technique. Thus leads to a second property of a screening method:

Property 2: The statistical screening technique should be flexible with respect to the objective function and constraints that can be handled.

The following objection to the LDF is a practical one. After the LDF has been constructed, the user must apply it to each new case to ascertain probable population membership. In the absence of computer support, it may be difficult to compute the score, especially if a large number of variables are involved. In cases where quadratic discrimination must be used, the objection is even more relevant. Also, the ultimate users may be concerned about the meaning of the output, e.g., they may wonder why a person's age category is multiplied by 2, and added to marital status multiplied by 3, etc. Getting the user to implement such an output may be problematic. Thus, a third property of a screening technique is:

Property 3: The form of the final output of a statistical screening technique should be meaningful for and amenable to practical use in screening new cases.

The LDF is constructed by using the MLE's of $\vec{U}_1$, $\vec{U}_2$ and Σ. For small sample sizes and a relatively large number of variables, the accuracy of these estimates may be poor. For example, Goldstein and Dillon (1977, pp. 7-10) compare the LDF constructed from 70% of a sample (325 cases) with the LDF constructed from the remaining 30% (130 cases). They found a great deal of instability in the relative rankings of variables, changes in coefficients from positive to negative, and coefficient signs that were not expected based on prior knowledge. The authors conclude that the LDF is highly sensitive to sample size, particularly when the variables are qualitative in nature as they were for this example.

Sample size is even more of a factor if there are departures from normality. Both Moore (1973) and Gessaman and Gessaman (1977), in comparing various discriminant analysis procedures, found that the LDF performed uniformly worse than the other procedures for instances where the assumptions did not hold and where the sample size was too small for asymptotic normality. In comparisons carried out by Gilbert (1968), the LDF performed quite favorably. However, Moore claimed that Gilbert's results were unrealistic since she assumed an underlying linear model for

which the problems of category ordering ("reversals" in the likelihood function as shown in the birth weight example) would not occur. Thus, the LDF should be restricted to instances where the sample size is large (a rule of thumb is that there should be at least 50 observations per variable).

Another desirable property of a statistical screening technique is that it should work well with the small sample size often encountered in practice:

Property 4: The statistical screening technique should be applicable for a wide range of sample sizes.

Several authors have addressed the issue of bias in the apparent error rate (sample probability of misclassification) in discriminant analysis. As discussed before, the expected apparent error rate is a negatively biased estimate of the expected actual error rate. However, this result is due to the Ex Post nature of the analysis and, as such, will tend to occur for any technique that develops decision rules based on sample data. Thus, this criticism cannot be leveled at the LDF alone.

Finally, the LDF is, by construction, an additive model. Thus, the approach may work poorly when there is an interaction effect among variables. The following example illustrates the problem:

Consider two binary variables X_1 and X_2 with the following sample results:

	X_1		
	0	1	Total
Π_1	100	100	200
Π_2	500	500	1000
Total	600	600	1200

	X_2		
	0	1	Total
Π_1	100	100	200
Π_2	500	500	1000
	600	6000	1200

(X_1, X_2)

	(0,0)	(0,1)	(1,0)	(1,1)	Total
Π_1	0	100	100	0	200
Π_2	500	0	0	500	1000
Total	500	100	100	500	1200

Note that $\hat{P}(\Pi_1|X_1) = \hat{P}(\Pi_1|X_2) = \hat{P}(\Pi_1)$
$\hat{P}(\Pi_2|X_2) = \hat{P}(\Pi_2|X_2) = \hat{P}(\Pi_2)$

Thus, X_1 and X_2 are both independent of $\Pi = <\Pi_1, \Pi_2>$. That is, by themselves X_1 and X_2 would not be selected as predictors. However, if we choose $D = <D_1, D_2>$ such that

$$D_1 = \{(0,0), (1,1)\}$$
$$D_2 = \{(0,1), (1,0)\}$$

then the probability of misclassification is zero. As shown earlier, however, the LDF cannot produce decision rule D.

In some applications, users attempt to get around this problem by introducing new variables X_1X_2, X_1X_3, $X_1X_2X_3$ etc. However, this quickly increases the number of coefficients to be estimated, particularly if the "fully saturated" model with all possible interactions is considered. Decisions about which interactions to include (e.g., only pairwise) are often made *a priori* with a resultant loss of information. In sum, approaches that can handle combined effects of variables should be preferred, as stated in property 5:

Property 5: The screening technique should be able to handle interaction effects among variables.

In summary, we have shown that the LDF approach may be inappropriate for screening with qualitative variables and, we have defined five properties that a screening technique should have. In the remaining sections of this Chapter, we review some of the other well known techniques for screening and critique them with respect to the five properties.

3.2 Multiple Regression Analysis

Multiple regression analysis may be used for screening by defining the dependent variable:

$$Y = \begin{cases} 1 \text{ if } \vec{x} \in \Pi_1 \\ 0 \text{ if } \vec{x} \in \Pi_2 \end{cases}$$

and estimating the equation

$$Y = B_0 + \sum_{i=1}^{k} B_i X_i \qquad (1)$$

by

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i \qquad (2)$$

in the usual way.

The decision rule $D = \langle D_1, D_2 \rangle$ is given by

$$D_1 = \left\{ \vec{x} \mid \hat{Y}_{\vec{x}} \geq 0.5 \right\}$$

$$D_2 = \left\{ \vec{x} \mid \hat{Y}_{\vec{x}} < 0.5 \right\}$$

The estimates $\hat{Y}_{\vec{x}}$ are interpreted as the probability that $\vec{x} \in \Pi_1$.

Because the regression equation is of a similar form to the LDF, this approach shares the problems associated with the LDF:

Property 1: The regression equation is affected by the ordering of categories unless all variable values are converted to indicator (dummy variables), e.g., if X_1 takes on four values X_{11}, X_{12}, X_{13}, X_{14}, new variables X_{11}, X_{12} and X_{13} are entered where

$$X_{1i} = \begin{cases} 1 \text{ if } X_1 = X_{1i} \\ 0 \text{ otherwise.} \end{cases}$$

However, this greatly increases the parameter space, e.g., if there are 10 variables each taking on 5 values, there are 41 coefficients to estimate in equation (1).

Property 2: The regression approach minimizes the sum of squares, i.e., $\sum (Y - \hat{Y})^2$. As shown in Chapter II, this is equivalent to minimizing the sample probability of misclassification. The approach does not handle the objective function ∇. Also, if the result is constrained to have $P \leq P^*$, users of the regression

approach merely shift the cut-off probability. Again, this assumes a monotonic relationship between probability as measured by $\hat{Y}$ and the actual likelihood of membership in Π_1. This assumption need not hold in general and will rarely hold for variables measured on a nominal scale.

Property 3: As with the LDF, the regression equation value can be difficult to compute for new cases and the equation itself (and coefficients) may have no meaning in the context of the problem.

Property 4: The regression equation requires fairly large sample sizes if the number of variables under consideration is high, and particularly if all the variables are converted to dummy variables as shown above. Also, the equation is sensitive to high correlation among variables, creating the problem of multicollinearity wherein the standard error of coefficient estimates may be very high.

Property 5: The regression equation can handle interaction effects but with the same problems of parameter proliferation as discussed for the LDF.

In addition to the above, there are some specific problems associated with the regression analysis approach for a binary dependent variable - see, for example, Hanushek and Jackson (1977). First, the assumption of homoscedasticity of error terms is violated as shown below:

If we rewrite (1) in matrix form as

$$Y_t = X_t B + e_t \qquad (3)$$

since $Y_t = 0$ or 1, we have

$$e_t = -X_t B \text{ or } 1 - X_t B.$$

Since $E(e_t) = 0$, we must have

$$\begin{aligned} E(e_t) &= -X_t B\, P(e_t = -X_t B) + (1-X_t B).P(e_t = 1-X_t B) \\ &= -X_t B\,(1-P(e_t=1-X_t B)) + (1-X_t B)P(e_t=1-X_t B) \\ &= -X_t B + P(e_t = 1-X_t B) \\ &= 0 \Longrightarrow P(e_t = 1-X_t B) = +X_t B \\ &\text{and } P(e_t = X_t B) = 1-X_t B \end{aligned}$$

but $\quad \text{Var}\ (e_t) = E\ (e_t^2) = X_t B\ (1 - X_t B) \qquad (4)$

which depends upon the observations.

Secondly, the regression equation (2) can yield estimates outside the range (0, 1) which is inconsistent with the probability interpretation attached to the estimates. Also, by the discussion above, $\hat{\text{Var}}\ (\hat{Y}_t) = \hat{Y}_t\ (1-\hat{Y}_t)$ which is negative for values outside the range (0, 1). Some users treat all estimated values less than 0 as 0, and values above 1 as 1. This yields an estimated variance of zero which is, clearly, a negatively biased estimate.

In sum, the regression approach suffers from the same problems as the LDF with respect to the five properties and has, in addition, important technical problems when dealing with a binary dependent variable.

3.3 Logit/Probit Analysis

Logit and probit analysis involve the specification of a functional form for the probabilities of class membership. In particular, the logistic function is

$$P(\Pi_1) = 1/\left(1+e^{-XB}\right) \qquad (1)$$

with, correspondingly,

$$P(\Pi_2) = 1/\left(1+e^{XB}\right) \qquad (2)$$

$$\text{then } \log \frac{P(\Pi_1)}{1-P(\Pi_1)} = \log P(\Pi_1) - \log\ [1-P(\Pi_1)] = X\vec{B}. \qquad (3)$$

Thus, the log of the ratio of probabilities is linear in the independent variables X.

The probit model is

$$P(\Pi_1) = \frac{1}{(2\Pi)^{\frac{1}{2}}} \int_{-\infty}^{XB} e^{-T^2/2} dT \qquad (4)$$

where T is N (0, 1). Clearly the probability increases monotonically with XB.

To estimate the coefficients in (1) and (2), we can use a MLE approach. If the observations $Y_1, \ldots, Y_n$ are ordered so that the n_1 cases from Π_1 appear first, then the cases from Π_2, we have:

$$L = P(Y_1, \ldots, Y_n) = \prod_{i=1}^{n_1} P_i \prod_{i=n_1+1}^{n} (1 - P_i)$$

$$= \prod_{i=1}^{n_1} P_i^{Y_i} (1 - P_i)^{1-Y_i} \qquad (5)$$

$$\text{and } \log L = \sum_{i=1}^{n_1} Y_i \log P_i + \sum_{i=n_1+1}^{n} (1-Y_i) \log (1-P_i) \qquad (6)$$

Substituting (1) for P_i in equation (6) and taking partial derivatives with respect to the B_k, i=1, ..., k, and setting them equal to zero, a set of k equations result which can be solved for the estimated coefficients b_k. In practice, the equations are difficult to solve. A similar approach may be followed for probit analysis.

I now discuss the logit/probit models in terms of the five properties previously developed.

Property 1: Invariance with respect to reordering of categories.

Both models are monotonic in XB. As such, similar problems occur with respect to the ordering of categories with cases like the example of 3.1, i.e., (0, 0), (1, 1) are members of Π_1 and (0, 1), (1, 0) are members of Π_2. The problem arises again because of the application of multiplication and addition functions to qualitative variables where, of course, such functions have no meaning.

Property 2: Flexibility with respect to objective functions.

With the MLE approach, the intent is to develop the best estimates of Y given the functional form assumed. This objective function is not among those discussed in Chapter II. Certainly, it would not be possible to maximize ∇, say. However, it is possible, as with the LDF and Regression Analysis, to set the cut-off probability in order to achieve a desired value of P*. That is, let c be the probability value such that at most nP* of the P_i are below it and at least n (1-P*) are above it. Then $D^* = \langle D_1^*, D_2^* \rangle$ is the decision rule with

$$D_1^* = \left\{ \vec{x} \mid P_{\vec{x}} \geq c \right\}$$

$$\text{where } P_{\vec{x}} = 1/\left(1+e^{-\vec{x}\vec{b}}\right)$$

is the estimated probability associated with $\vec{x}$.

As discussed before, this is not necessarily an optimum procedure since D_1^* was not selected as best among all possible decision rules with $P \leq P^*$.

Property 3: Amenable to practical use.

Just as with the LDF and the regression analysis approach, the probabilities are difficult to compute by hand and the form of the probability equation (1) is difficult for the average user to understand or appreciate.

Property 4: Applicable for a wide range of sample sizes.

The sample sizes needed for logit/probit analysis are similar to those required for the LDF or regression analysis approach since the same number of coefficients must be estimated from the sample data. Thus, for most problems with 10 or more variables, sample sizes should be in excess of 500.

Property 5: Ability to handle interaction effects

Consider the derivative of (1) with respect to the variable X_i:

$$\frac{\partial P(\Pi_1)}{\partial X_i} = \frac{\partial \left(1/1+e^{-XB}\right)}{\partial X_i} = \frac{1}{\left(1+e^{-XB}\right)^2} \frac{\partial}{\partial X_i}\left(e^{-XB}\right)$$

$$= \frac{1}{\left(1+e^{-XB}\right)^2} B_i e^{-XB}$$

$$= \frac{B_i}{\left(1+e^{-XB}\right)} \frac{e^{-XB}}{1+e^{-XB}} =$$

$$= \frac{B_i}{\left(1+e^{-XB}\right)} \frac{e^{-XB}e^{XB}}{e^{XB}+1} = B_i P(\Pi_1) \frac{1}{1+e^{XB}}$$

$$= B_i P(\Pi_1) (1-P(\Pi_1)) \qquad (7)$$

Expression (7) shows that the logit model handles interactions since the value of the derivative is a function of B_i and P which is itself a function of all the independent variables.

Similarly, for the probit function,

$$\frac{\partial P}{\partial X_i} = \frac{\partial}{\partial X_i} \frac{1}{\sqrt{2\Pi}} \int_{-\alpha}^{XB} e^{-T^2/2}\, dT$$

$= B_i \phi(XB)$ where $\phi(XB)$ is the value of the standard normal density at the point XB.

In sum, logit and probit analysis offer some advantages over the LDF and regression analysis for the type of problem under consideration. Nonetheless, they assume an underlying model form which may or may not be appropriate in a particular instance. This is, therefore, an undesirable feature if alternative procedures are available which do not involve such an assumption.

3.4 The Multinomial Model

If a random sample of size n is selected from $\Pi = \langle \Pi_1, \Pi_2 \rangle$, then the estimated probability that $\vec{X} = \vec{x}$ in population Π_1 is

$$\hat{P}(\vec{X} = \vec{x} | \Pi_1) = \frac{n_1(\vec{x})}{n_1} \tag{1}$$

where $n_1(\vec{x})$ is the number of sample observations in Π_1 with the vector $\vec{x}$.

Similarly

$$\hat{P}(\vec{X} = \vec{x} | \Pi_2) = \frac{n_2(\vec{x})}{n_2} . \tag{2}$$

Thus
$$\hat{P}(\Pi_1 | \vec{X} = \vec{x}) = \frac{\hat{P}(\Pi_1 \text{ and } \vec{X} = \vec{x})}{P(\vec{X} = \vec{x})}$$

$$= \frac{\hat{P}(\Pi_1)\ \hat{P}(\vec{X} = \vec{x} | \Pi_1)}{P(\vec{X} = \vec{x})}$$

$$= \frac{\frac{n_1}{n} \cdot \frac{n_1(\vec{x})}{n_1}}{\frac{n_1(\vec{x}) + n_2(\vec{x})}{n}} \tag{3}$$

$$= \frac{n_1(\vec{x})}{n_1(\vec{x}) + n_2(\vec{x})}$$

and
$$\hat{P}(\Pi_2 | \vec{X} = \vec{x}) = \frac{n_2(\vec{x})}{n_1(\vec{x}) + n_2(\vec{x})} \tag{4}$$

The intuitive decision rule based on (3) and (4) is to classify the case into Π_1 if

$$\hat{P}(\Pi_1 | \vec{X} = \vec{x}) > \hat{P}(\Pi_2 | \vec{X} = \vec{x}),$$

into Π_2 otherwise (randomly assigned if equality holds). That is,

$$\hat{D} = \langle \hat{D}_1, \hat{D}_2 \rangle$$

is such that

$$\hat{D}_1 = \left\{ \vec{x} | n_1(\vec{x}) > n_2(\vec{x}) \right\} \quad \hat{D}_2 = \left\{ \vec{x} \mid n_1(\vec{x}) < n_2(\vec{x}) \right\}.$$

While this "full multinomial" approach has some appealing features—particularly the simplicity of the decision rule—it has one overwhelming disadvantage: for a problem with k variables each taking on s_i values, there are

$$m = \prod_{i=1}^{k} s_i$$

possible observation vectors, e.g., for k = 5 and $s_i = 5 \ \forall_i$, $m = 5^5 = 3125$. For a typical sample, many of the states will not appear at all, some will appear in Π_1 but not Π_2 and vice versa, while others will appear so infrequently as to make the estimates highly unreliable. If a particular realization $\vec{x}$ appears once in Π_1 and never in Π_2, $P(\Pi_1 | \vec{X} = \vec{x}) = 1$. The same probability would result if $\vec{x}$ appeared 50 times in Π_1 and never in Π_2, yet our intuitive faith in the probability estimate would be much higher for the latter finding. If the user is faced with the task of allocating a new case with an observation vector $\vec{x}$ not found in the sample, he has no mechanism for making the choice.

For these reasons, the full multinomial approach is considered unacceptable for most applications. The aim, instead, is to modify the approach to reduce the state space. For example, Hills (1967) defines a "nearest neighbour" rule for the case where the variables are all dichotomous. For a given observation vector $\vec{x}$ he defines a set of vectors that differ from $\vec{x}$ in no more than r positions, i.e.,

$$T_j = \left\{ \vec{y}_j | (\vec{x}-\vec{y}_j) \quad (\vec{x}-\vec{y}_j) \leq r \right\}. \qquad (5)$$

Then the rule for $\vec{x}$ is based on the vectors $\vec{y}_j$. That is, the decision rule is as follows:

$$\begin{aligned} &\vec{x} \in \Pi_1 \text{ if } \sum_{T_j} n_1(\vec{y}_j) > \sum_{T_j} n_2(\vec{y}_j) \\ &\vec{x} \in \Pi_2 \text{ if } \sum_{T_j} n_1(\vec{y}_j) < \sum_{T_j} n_2(\vec{y}_j) \qquad (6) \\ &\text{(randomly allocate if equality holds).} \end{aligned}$$

Several other models have been developed for the case of dichotomous variables, such as the Bahadur model in which, for $\vec{X} = (X_1, \ldots, X_m)$

$$P_{ij} = E_i (X_j), \quad i = 1, 2$$

$$Z_{ij} = \frac{X_j - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}}$$

$$\rho_i(jk) = E (Z_{ij} Z_{ik})$$

$$\rho_i(12\ldots m) = E (Z_{i1} Z_{i2} \ldots Z_{im})$$

Then $$P (\vec{x} | \Pi_i) = \prod_{j=1}^{mk} P_{ij}^{x_j} (1 - P_{ij})^{1-x_j} \times$$

$$\times \left(1 + \sum_{j<k} \rho_i (jk)\, Z_{ij} Z_{ik} + \sum_{j<k<l} \rho_i (jkl)\, Z_{ij} Z_{ik} Z_{il} + \ldots + \rho_i (1, 2, \ldots, m)\, Z_{i1} Z_{i2} \ldots Z_{im}\right) \qquad (7)$$

Then, the classification procedure is the likelihood ratio

$$\frac{P(\vec{x} | \Pi_1)}{P(\vec{x} | \Pi_2)}$$

with estimates of all the above parameters substituted where appropriate.

By assuming higher order correlations are zero, the expression (7) can be reduced significantly.

Another approach developed by Martin and Bradley (1972), involves a description of the density of $\vec{x}$ in terms of otthogonal polynomials. Matsuita (1954, 55, 57) develops classification rules based on placing an observation into the class such that the estimated distributional distance is maximized.

The rules discussed in this section are briefly reviewed below in terms of the properties introduced in Section 3.1.

Property 1: Invariance w.r.t. categories of variables

For those procedures assuming dichotomous variables, this property is not relevant. However, the assumption of dichotomous variables is rather restrictive for most practical applications.

Property 2: Flexibility w.r.t. objective functions

The procedures, with the exception of the ones based on distributional distance, are optimum under the condition that the assumed model holds—since the rules are likelihood ratio based. However, the procedures are not amenable to other objective functions or to control of the screening parameters P and Q. It appears that estimation of an assumed underlying distribution takes precedence over accurate prediction of class membership.

Property 3: Amenable to practical use

The full multinomial approach is simplest to apply but, as has been pointed out, there may be many instances when a new case cannot be classified because the observation vector $\vec{x}$ did not occur in the sample. The other procedures are all rather complicated involving, in the Bahadur model for example, computing expressions for $P(\vec{x}|\Pi_i)$, $i = 1, 2$, with a large number of parameter estimates (the P_{ij} and $\rho_i\ (jk)$).

Property 4: Applicable for a wide range of sample sizes

The multinomial approach clearly requires very large sample sizes in order to ensure sufficient observations for every state. Similarly, the other procedures require large sample sizes to ensure reliable estimates of the many parameters. Small sample sizes are adequate only in the instances when there are very few variables taking on few values and where no observation patterns are rare.

Property 5: Handles interactions among variables

The approaches do handle interactions in the model form but the parameter space tends to become unwieldy. This often forces users to make a priori decisions concerning the order of interaction terms to include.

3.5 Automatic Interaction Detection (AID)

In the Preface to their book Searching for Structure, Sonquist, Baker and Morgan say that they developed the AID technique "...in rebellion against the restrictive assumptions of conventional multivariate techniques...". Thus, AID represents an initial step towards an approach that is appropriate for qualitative independent variables.

The AID approach involves a repeated one-way analysis of variance technique to explain as much variance in the dependent

variable as possible. Although the approach was designed for a continuous dependent variable, the authors state that a dichotomous dependent variable may be used if (in our notation) $0.2 \leq E \leq 0.8$. Thus, AID can be used for statistical screening under these circumstances.

The error variance, with no knowledge of the dependent variables, is, as we have seen,

$$\Sigma (Y - \bar{Y})^2 = \Sigma Y^2 - N\bar{Y}^2 = E\,(1 - E) \qquad (1)$$

For any two groups formed by the data (i.e., a decision rule $D = \langle D_1, D_2 \rangle$), the error variance is

$$\Sigma Y_1^2 - N_1 Y_1^2 + \Sigma Y_2^2 - N_2 \bar{Y}_2^2$$

$$= \Sigma Y^2 - N_1 \bar{Y}_1^2 - N_2 \bar{Y}_2^2$$

$$= E + P - 2PQ \text{ as shown in Chapter II.} \qquad (2)$$

The net reduction in variance is (1) - (2), or

$$P - 2PQ - E^2. \qquad (3)$$

The AID algorithm searches through the k variables X_1, X_2, ..., X_k and, for each variable, examines the possible splits of the data using the categories of that variable. The variable for which (3) is maximized for specified categories is chosen for entry. This procedure is then applied to each "split" of the population so generated. No further splitting occurs if:

- the marginal reduction in variance is below a pre-specified threshold;
- one of the new groups would have a total number of cases below a specific threshold;
- the total number of splits has exceeded a pre-specified amount.

Bishop et al (1975) state that the AID process has the drawback that it does not take into account the sampling variability in the data. They cite a study by Einhorn (1972) which demonstrated that the AID algorithm consistently created apparent structures where none existed in reality. That is, the AID algorithm fits the data well but not necessarily the underlying situation.

We now discuss the AID algorithm in terms of the properties:

Property 1: Unaffected by random reordering of categories

Because the algorithm examines single characteristics at a time, it is unaffected by random reordering of categories. Thus, the algorithm satisfies property 1.

Property 2: Flexible w.r.t. objective functions

The approach is designed to maximize explained variance, although it may be possible to consider the same sequential splitting process with a different splitting function. However, there is little control over the parameters P and Q. Thus, the AID algorithm, in its current form at least, does not satisfy property 2.

Property 3: Useful output

The usefulness of the output depends upon the number of groups formed by the data, and the number of variable values in each group. If good results can be obtained while constraining the number of groups and group membership, the output can be meaningful. Otherwise, the results may represent merely a good description of the structure of the data. Nonetheless, the form of the output, linking characteristics by the logical and operator, is more meaningful than other techniques with output in the form of mathematical operations on qualitative variables. From this viewpoint, then, the approach satisfies property 3.

Property 4: Applicable for wide range of sample sizes

The AID algorithm works with both large and small sample sizes. However, the search type algorithm requires a large number of calculations to choose among alternative splits at each stage. Thus, it tends to be more effective with small sample sizes.

Property 5: Ability to handle interaction effects

Because the algorithm considers all categories of variables at each stage, it explicitly handles interaction effects. However, it does not handle all possible interaction effects because of the sequential nature of the process.

In sum, the AID algorithm comes closest to meeting the properties desired of an approach for screening with qualitative variables. However, it is somewhat restrictive. Nonetheless, it demonstrates the power of advances in data processing capabilities to provide more in-depth analysis of data. The aim,

then, is to develop more generalized procedures for analyzing qualitative data for screening populations. This is the subject of the last two chapters of this dissertation.

3.6 Summary

This Chapter provided an overview of the major competing statistical screening techniques. In particular, we discussed the applicability of each technique for practical problems involving variables measured at the nominal level. In general, the techniques failed because they were designed for higher level variables.

The criticism of the techniques led to the definition of five properties that should be sought after in development of a new technique for qualitative variables:

- Property 1: The statistical algorithm should not be affected by random reordering of the categories of each variable.
- Property 2: The technique should be flexible with respect to the objective function and constraints that can be handled.
- Property 3: The form of the final output of the approach should be meaningful for and amenable to practical use in screening new cases.
- Property 4: The statistical screening technique should be applicable for a wide range of sample sizes.
- Property 5: The screening technique should handle interaction effects among variables.

Of the techniques reviewed, the AID algorithm came closest to satisfying the above properties because it was developed specifically for qualitative independent variables. The aim of our research is to go beyond the AID algorithm to develop a new class of procedures that have wider applicability, which have greater power, and which more closely satisfy the above properties. Chapter IV describes our initial efforts to develop these procedures.

IV. THE MONTE CARLO APPROACH

IV. THE MONTE CARLO APPROACH

In this Chapter, we describe a new approach to screening based on Monte Carlo simulation. Results of an actual application of the methodology are also presented. To our knowledge, the Monte Carlo approach as developed here has not been used before for purposes of statistical screening.

4.1 Background

In Chapter III, we briefly reviewed the general types of screening techniques in common use. This review demonstrated that, for practical screening problems with qualitative variables and relatively small sample sizes, all of the approaches had some serious weaknesses. The Monte Carlo approach described here presents a first attempt to develop a new method that would be appropriate to this type of problem.

Thus, the approach taken was to start from the basic concepts of screening to develop a straightforward, easily understood approach that would compare favorably with more established techniques and, under certain conditions, outperform those techniques. The first step, then, was to consider trial and error selection.

4.2 Trial and Error Profile Selection

We introduce here the term "profile" to describe the set of characteristics associated with cases from D_1. That is, a new case fits the "profile" associated with a rule $D = \langle D_1, D_2 \rangle$ if the vector of characteristics, $\vec{x}$, is a member of D_1.

An intuitive approach to developing a good rule D might be based on trial and error selection of possible variables and variable values. One such approach is described below, focusing on the parameter estimates p_{ij} and q_{ij} for each variable value:

1. Examine the (p_{ij}, q_{ij}) combination for each variable value x_{ij}, $i=1, \ldots, k$, $j=1, \ldots s_i$.

2. Identify variable values with relatively high values of q_{ij}. Since these variable values generally have a low value of p_{ij}, combine these variable values with the or operator.

3. Identify other variables with relatively high values of $p_{ij} \cdot q_{ij}$. These variables can then be combined with the variables in 2. with an and operator.

This process is illustrated below with an actual example. Note that the initial aim was to develop high values of p.q rather than $\hat{V}$, for example, since this work preceded the development of more complex objective functions.

Table 4.1 shows the variables that were used (some preliminary work was done to develop this reduced list) ranked in order of individual q values.

TABLE 4.1

Variable #	q	p
1	.7500	.0264
2	.5405	.0488
3	.4651	.0567
4	.4595	.0488
5	.4565	.0607
6	.4474	.0501
7	.3968	.1662
8	.3875	.1055
9	.3818	.0726
10	.3654	.0686
11	.3377	.1016
12	.3367	.1293
13	.3294	.1121
14	.3290	.2045
15	.3141	.2519
16	.3095	.2216
17	.3076	.5488

The first trial profile was as follows: (1 or 2 or 3 or 4 or 5 or 6) and 17. The reason for this construction was that the first six variables had q values above 0.4474, with little difference among variables 3 through 6. Variable 7 was noticeably lower at 0.3968. Then, it was noted that the q value for variable 17 was not much lower than the others while p_{17} was by far the highest (alternatively, $p_{17} \cdot q_{17}$ was a maximum). Thus, variable 17 was included in an and combination.

This profile had a value for p of 0.1121 and of q = 0.5176, compared to the random rate e = 0.232.

Because p was lower than desired, the next three trial profiles involved adding variables 7, 8 and 9 in turn in an or combination as follows. The fifth profile was just the first six variables connected with or. The sixth through eighth profiles successively added variables 7, 8 and 9 in an or combination. Variable 17 was not included in any of these profiles, in order to achieve a higher q.

Results for these eight trial profiles are given in Table 4.2:

TABLE 4.2

Profile #	Profile Description	p	q	p*q
1	(1 or 2 or 3 or 4 or 5 or 6) and 17	.1121	.5176	.0580
2	(1 or 2 or 3 or 4 or 5 or 6 or 7) and 17	.1741	.4242	.0738
3	(1 or 2 or 3 or 4 or 5 or 6 or 7 or 8) and 17	.1834	.4101	.0752
4	(1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9) and 17	.2058	.4103	.0844
5	1 or 2 or 3 or 4 or 5 or 6	.1926	.4521	.0810
6	1 or 2 or 3 or 4 or 5 or 6 or 7	.2995	.3700	.1108
7	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8	.3193	.3554	.1134
8	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9	.3575	.3506	.1253

Exhibit 4.1 plots the results in terms of the profile diagram of Chapter II. The circled numbers represent the profiles; the other numbers are the original variables. The exhibit shows that the trial profiles generally outperform the single variables.

A definition of inadmissibility is as follows: a profile (p, q) is inadmissible if $\exists$ profile (p*, q*) such that $p^* \geq p$ and $q^* \geq q$ and either $p^* > p$ or $q^* > q$. That is, the profile (p*, q*) dominates (p, q). In this context, variables 3, 4, 5, 6, 8, 9, 10, 11 are dominated by trial profile ①. Variables 12, and 13 are dominated by profile ⑤. Variables 14, 15 and 16 are dominated by profile ⑥. Also, profiles ② and ③ are dominated by ⑤. The potentially admissible profiles (or variables) are 1, 2, 17 and ①, ⑤, ④, ⑥, ⑦ and ⑧. Note that there may exist profiles not yet discovered by the analysis which dominate the above profiles.

Exhibit 4.1 also illustrates the tendency for profiles to cluster. For example ②, ③, ④ and ⑤ are fairly similar as are ⑥, ⑦ and ⑧. Variables 3, 4, 5 and 6 are alike also. This is important in practice for it provides the user some flexibility in choosing the final solution.

Overall, however, the gain in going from single variables to multi-variable profiles appears to be small in this application. This suggests that the trial and error approach is rather poor as a guide to combining variables. The reason, of course, is that the approach does not use any information covering the interrelationships among the variables.

There are two general directions one can take to improve this situation:

1) Generate a much larger set of profiles from which the best may be selected.

2) Develop techniques that make use of information on the relationship between two variables.

In the remainder of this Chapter, we pursue the first direction. Chapter V introduces new techniques of the second type.

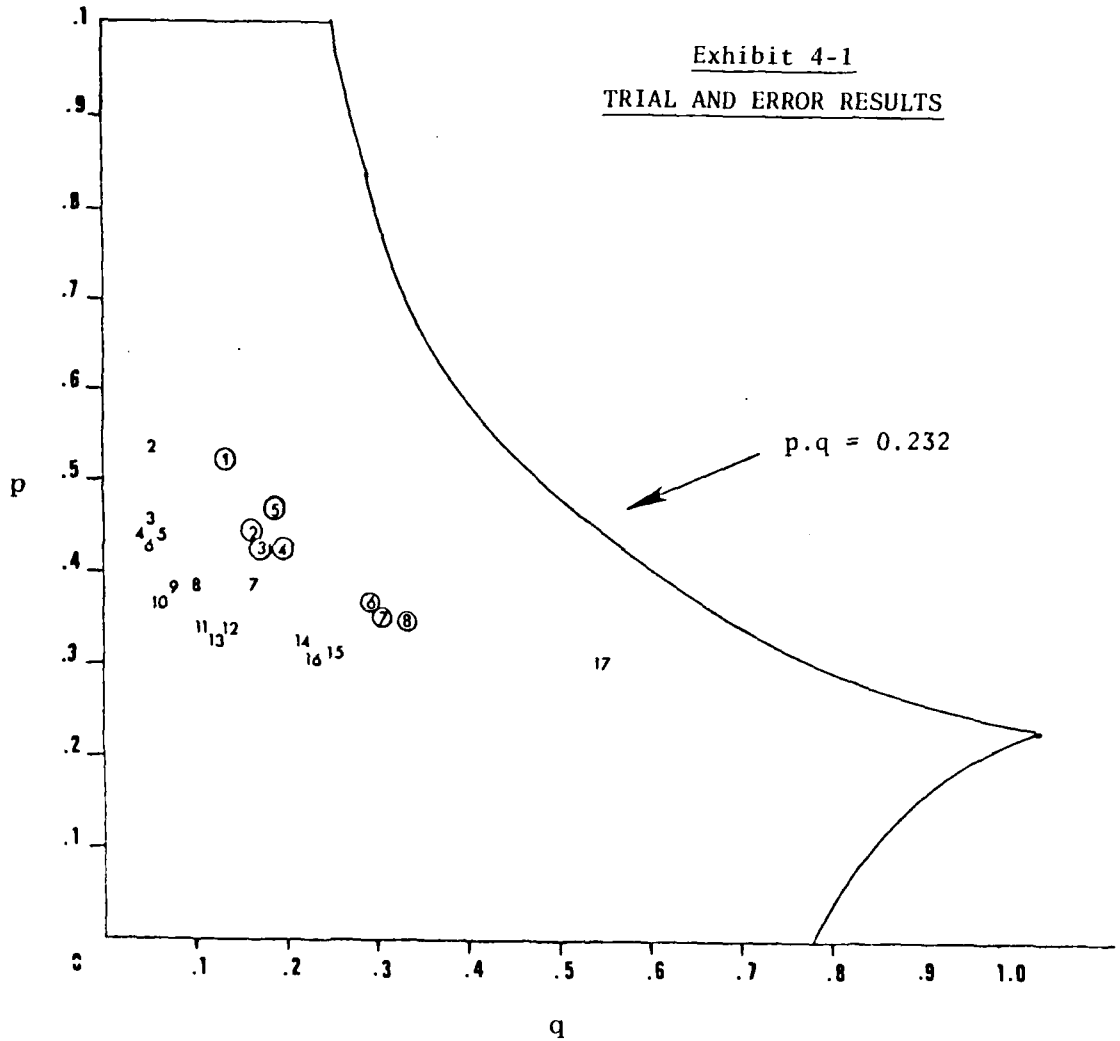
4.3 Monte Carlo Approach

The Monte Carlo approach is, in effect, a trial and error approach without human intervention. The computer is instructed to pick variable values at random and combine them in a random fashion to form a profile. On repeated trials, the hope is that a profile will be found which is sufficiently close to optimal for practical purposes.

Theoretically, the process is as follows: randomly select the initial x_{ij}, say, from the $\prod_{i=1}^{k} s_i$

Exhibit 4-1

TRIAL AND ERROR RESULTS



possible variable values. Randomly select an operator from the set (and, or, not). Then, randomly select a variable from the remaining set and continue until a random number of variable values have been included. For each profile so generated compute p,q, p.q and ∇ (or any other objective function of interest). Select the profile that maximizes the objective function.

Note that this procedure does not make use of any information about the interrelationship of variables. Instead it relies on the power of repeated trials to locate the interrelationships that have strong discriminatory power.

Example

This example is again taken from the study of Medicaid cases. The actual procedure followed differed somewhat from the theoretical description above. Variable 17 was pre-specified to be in the profile because it had such a high value of p. Also variable 1 was included because it had a particularly high value of q. That is, the form was specified to be as follows:

> 17 and (1 or V_1 or V_2 or ... or V_m) where V_1 to V_m were to be picked at random from the other 15 variables x_2 to x_{15}. 7 to 10 variables were to be added for each trial, with the actual number selected being a random variable ($P(7) = P(8) = P(9) = P(10) = \frac{1}{4}$).

Clearly, this represents a rather constrained example of the Monte Carlo approach. By constraining the solution, the number of very poor profiles generated is reduced but there is a corresponding loss of power.

Table 4.3 shows the results for the nine profiles judged best from the 200 profiles generated in this fashion.

TABLE 4.3

Monte Carlo Profile Number	Profile Description	p	q	p·q
[1]	17 and (1,4,5,6,9,11,12, 14,$\overline{15}$,16)	.3391	.3658	.1240
[2]	17 and (1,2,6,8,10,11,12, 14,$\overline{15}$)	.3061	.3707	.1135
[3]	17 and (1,3,4,6,9,11,13, 15,$\overline{16}$)	.2995	.3833	.1148
[4]	17 and (1,2,3,4,9,10,12, 13,$\overline{16}$)	.2982	.3850	.1148
[5]	17 and (1,3,5,6,7,9,10,11, 13,$\overline{16}$)	.2810	.4038	.1135
[6]	17 and (1,2,4,6,8,9,12, 13,$\overline{16}$)	.2757	.4067	.1121
[7]	17 and (1,3,4,5,8,9,10, 11,$\overline{12}$,16)	.2704	.4000	.1082
[8]	17 and (1,2,3,4,9,10,12, 16)	.2573	.4256	.1095
[9]	17 and (1,2,4,5,9,10,12, 16)	.2559	.4222	.1052

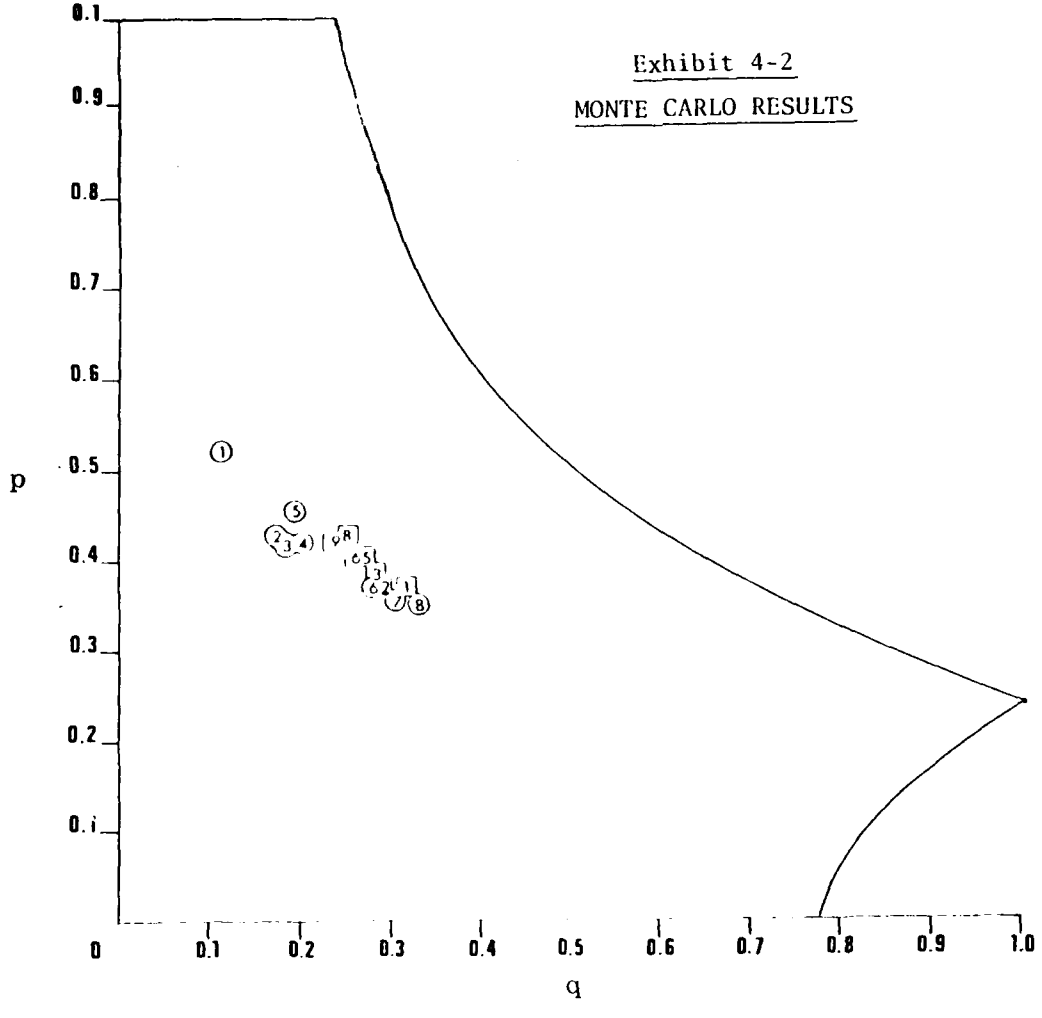
To show the improvement achieved with the random selection, we compare the Monte Carlo profiles with the trial and error profiles:

- profile ⑥ is dominated by profile [3]
- profiles ③, ④ and ② are dominated by profile [8]
- profile ⑦ is dominated by profile [1]

More importantly, perhaps, no trial and error profile dominates a Monte Carlo profile. Exhibit 4.2 shows the Monte Carlo

Exhibit 4-2

MONTE CARLO RESULTS



profiles [4 and 7 are not shown since the results are so similar to 3 and 6 respectively]. Some of the trial and error profiles are shown for comparative purposes.

Of course, these results are not unexpected since these are the best 9 out of 200 Monte Carlo trials, whereas there were only 8 trial and error profiles. Nonetheless, it does demonstrate the general assumption underlying the Monte Carlo approach: the greater the number of trials, the greater the chance of finding a good profile.

In examining Table 4.3, alternative ways of improving the results may be considered. For example, since the 9 profiles shown there are the "best" (at least in terms of p.q), we could identify common characteristics of these profiles. Table 4.4 displays the results by variable for the 9 profiles.

TABLE 4.4

	VARIABLES															
Profile	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	X			X	X	X			X		X	X		X	X	X
2	X	X				X		X		X	X	X		X	X	
3	X		X	X		X			X		X		X		X	X
4	X	X	X	X					X	X		X	X			X
5	X		X		X	X	X		X	X	X		X			X
6	X	X		X		X		X	X			X	X			X
7	X		X	X	X			X	X	X	X	X	X			X
8	X	X	X	X					X	X		X				X
9	X	X		X	X				X	X		X				X
Total	9	5	5	7	4	5	1	3	8	6	5	7	4	2	3	8

The results show that variables 9 and 16 appear in 8 out of the 9 profiles and always together. Variable 7, on the other hand, appears only once, and variable 14 only twice. This suggests that new profiles could be generated using a smaller set of variables. For example, restricting the selection to only

those variables appearing 6 or more times, we obtain the reduced set [1, 4, 9, 10, 12, 16] .

One of the advantages of the Monte Carlo approach is the information provided by the profiles randomly generated. It is much easier to examine the best profiles developed to determine clues as to effective combinations of variables. In other words, the Monte Carlo approach can be linked to an approach involving human intervention, taking advantage of the computer's power to identify promising profiles.

4.4 Evaluation

The Monte Carlo approach is, of course, Ex Post. Thus, there is always the danger that the "best" Monte Carlo profile merely fits the sample data rather than reality. Two methods may be adopted to test the results:

1) Develop the profile(s) on a subset of the data base, then test them on the remainder of the data base.

2) Develop the profile(s) on one data base and then test them on another sample from the same population.

In this section, we discuss an empirical test of the Monte Carlo approach based on the second method. Again returning to the New Hampshire Medicaid example, profiles were developed based on a sample taken from four District Offices. These profiles were then applied to a later sample of Medicaid cases drawn from the same four offices. Using these data, we compare the actual versus predicted performance.

Table 4.5 shows the values of p,q and $\hat{\nabla}$ (denoted p_B, q_B, $\hat{\nabla}_B$) developed from the original sample and resulting values obtained when the profiles were applied to a new sample of cases. Separate profiles were used for the four District Offices for each of two population types: AI = Adult Independent and NH = Nursing Home. The value of e was assumed to have remained the same.

TABLE 4.5

		Test Results							
		n_A	p_A	q_A	e	p_B	q_B	$\hat{\nabla}_A$	$\hat{\nabla}_B$
Manchester	NH	369	.276	.559	.355	.346	.607	.259	.383
	AFDC	137	.139	.579	.421	.302	.538	.099	.151
Concord	NH	194	.387	.627	.303	.346	.607	.551	.479
	AFDC	64	.125	.875	.421	.302	.538	.258	.151
Berlin	NH	144	.111	.625	.176	.124	.656	.402	.464
	AFDC	50	.060	1.000	.375	.093	.750	.192	.175
Conway	NH	36	.139	.800	.125	.125	.656	.818	.607
	AFDC	28	.179	.600	.375	.093	.750	.192	.175

Comparison of P Values

Treating p as a sample proportion, it is possible to test the hypothesis:

$$H_0: \ P_A = P_B = P \text{ vs. } H_A: \ P_A \neq P_B.$$

That is, reject H_0 at the 5% level if

$$P_A < P_B - 1.645 \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n} + \frac{1}{m}\right)}$$

where $\bar{p}$ is the pooled estimate of P, and n and m are the respective sample sizes.

Applying this test to the sample data, H_0 could not be rejected for any of the eight tests. In other words, there is no evidence to support the contention that the proportion of cases fitting the profile is less than expected.

Comparison of Q Values

Treating Q as a binomial proportion within the reduced set of cases fitting the profile, we can perform a similar test of the hypothesis:

$$H_0: \ Q_A = Q_B = Q \quad \text{vs.} \quad H_A: \ Q_A < Q_B$$

Reject H_0 at the 5% level if

$$q_A < q_B - 1.645 \sqrt{q(1-q)\left(\frac{1}{n^*_A} + \frac{1}{n^*_B}\right)}$$

where q is the pooled estimate of Q, and n^*_A, n^*_B are the number of sample cases fitting the profile in each sample.

Again, H_0 could not be rejected for any of the eight tests. However, for two of the tests the effective sample sizes were too low for a reasonable test. Nonetheless, 5 of the 8 q_A values were above the corresponding q_B values. Thus, the profile performed about as expected with respect to the parameter q.

Comparison of ∇ Values

Although the profiles were not developed to maximize ∇, it is instructive to compare the realized values of ∇ in the two samples.

First, it is interesting to note that $\hat{\nabla}$ appears to be inversely correlated with the value of e. That is, the lower the observed value of e, the greater the absolute value of ∇. This is intuitively reasonable since there is greater potential to increase q relative to e when e is small. The value of $\hat{\nabla}$ is, of course, sensitive to the difference q-e.

Second, the observed values $\hat{\nabla}_B$ are greater than their counterparts $\hat{\nabla}_A$ in 5 of the 8 comparisons. This is a surprising finding, given that we would expect the estimates $\hat{\nabla}_B$ to be optimistically biased given that the profiles were developed Ex Post on relatively small sample sizes.

Because the sample sizes for each office in the original sample were not available, it was not possible to test the hypotheses $\nabla_A = \nabla_B$. Furthermore, since the value of e was assumed to be the same in each sample, the tests would not be accurate.

Comparison with Other Results

Because other techniques have been used for the same problem of detecting ineligible Medicaid cases, it is instructive to compare their performance with the results here.

For example, South Carolina and the District of Columbia have used discriminant analysis with the results as shown in Table 4.6.

TABLE 4.6

	p	q	e	$\hat{\nabla}$
South Carolina	0.2	0.38	0.19	0.242
District of Columbia (1)	0.2	0.18	0.13	0.072
District of Columbia (2)	0.2	0.45	0.24	0.244
District of Columbia (3)	0.2	0.14	0.07	0.116

The three sets of results for D.C. refer to three different types of eligibility error types. The results shown are the results based upon the sample used to develop the discriminant function. Therefore, the actual results may be different. To compare these results with the Monte Carlo results, we examine instances where the values of e are similar:

- For values of e of 0.176 and 0.125, the Monte Carlo results (as shown in Table 4.5) were values of $\hat{\nabla}_B$ of 0.464 and 0.607 respectively; whereas, for discriminant analysis, values of e of 0.13 and 0.07 were associated with $\hat{\nabla}$ values of only 0.072 and 0.116 respectively.

- For values of e of 0.355 and 0.303, the Monte Carlo achieved $\hat{\nabla}_B$ values of 0.383 and 0.479; whereas, for discriminant analysis, values of e of 0.19 and 0.24 had $\hat{\nabla}$ values of 0.242 and 0.244.

Because these results were based on samples from different States, definitive conclusions cannot be made. Nonetheless, it appears that the Monte Carlo approach outperforms the discriminant function approach.

The Social Security Administration uses the AID algorithm to help identify ineligible Supplemental Security Income cases. They achieved the following results:

$$p = 0.11$$
$$q = 0.57$$
$$e = 0.2$$
$$\hat{\nabla} = 0.33$$

The value of $\hat{\nabla}$ is greater than the values obtained by the discriminant analysis approach for similar values of e. This is not unexpected since, as we pointed out in Chapter III, the AID algorithm came closest to satisfying the desired properties of a screening procedure. Furthermore, the Monte Carlo approach appears to outperform the AID algorithm indicating that it may be a viable alternative. In Chapter V, we introduce further refinements to the Monte Carlo approach which perform even better than the figures presented here.

We should also point out that the Monte Carlo profiles were not selected to maximize $\hat{\nabla}$. Thus, we would expect that, if maximization of $\hat{\nabla}$ had been the specified objective function, the resulting $\hat{\nabla}$ values would have been higher.

These comparisons are not rigorous. A more rigorous comparison would require the application of each technique to the same data base. Because of the limited resources for programming, we decided to defer testing of the algorithms until the new methods presented in Chapter V could be programmed. As will be seen, the programming task is quite extensive.

4.5 Summary

In this Chapter, we have presented a new type of statistical screening approach that takes advantage of the power of the computer to identify effective combinations of variables. The approach involves no assumptions about the underlying distribution of variables. It classifies every observation into one of the two populations. Despite the fact that the methodology was developed from an intuitive basis, without any closed-form analysis to demonstrate the "optimality" of the approach, the empirical results suggest that the approach is highly effective. In particular we have shown that the sample-based results were duplicated in an actual test of the profiles on a new sample.

The approach also satisfies the properties introduced in Chapter III, namely:

- Property 1: The methodology, by construction, is clearly unaffected by changes in the ordering of categories of each variable since the approach randomly picks variable values from the available ones.

- Property 2: The approach can handle any objective function that is based on sample statistics since it automatically calculates the objective function for each random profile

developed. Constraints can be handled easily by choosing the best profile among those satisfying the constraint.

- Property 3: The results are easy to use and interpret. The solution is in the form of a sequence of variables linked by and or or. The user need only check the characteristics of each new case against the logical expression. No mathematical computations are required.

- Property 4: The approach can handle any sample sizes, although the advantages of the technique over other techniques such as the LDF tend to diminish as sample size increases. For example, the LDF's assumption of multivariate normality becomes less problematic. For small sample sizes, the technique has clear advantages over approaches based on the multinomial model which suffers from state sparseness.

- Property 5: The approach handles interactions explicitly since it seeks out effective combinations of variable values.

Despite the strengths of the Monte Carlo approach, we believe it barely scratches the surface of what appears to be a new direction for screening problems with qualitative variables. In Chapter V, we develop the initial concepts embodied by the Monte Carlo approach into a much wider class of procedures.

V. NEW SCREENING PROCEDURES

V. NEW SCREENING PROCEDURES

5.1 Introduction

In Chapter II, we provided some background to the screening problem, including an overview of the various objective functions that could be relevant. In Chapter III, we reviewed some of the established screening techniques and demonstrated that none of these techniques were entirely satisfactory under certain conditions. We also suggested five properties that a screening technique should have. In Chapter IV, we presented an initial effort to apply a new technique that was designed specifically for the problem under consideration. Results of that application were also presented. In this chapter, we build upon the work of those chapters in order to propose a new group of screening procedures. First, we consider sequential algorithms for general and specific objective functions. Then, we develop procedures with a pre-specified form of output.

5.2 The General Sequential Algorithm

As discussed in Chapter III, it is important that a screening algorithm be amenable to different objective functions. Also, in Chapter II, we showed that most objective functions of interest are well-behaved functions of the parameters P, Q and E. Thus, we can assume a general objective function that is to be maximized (without loss of generality):

$$O = f(P, Q, E). \tag{1}$$

Since we are dealing with sample data, the aim of the algorithm is to find the maximum of the sample value of O, that is to find

$$\hat{O}^* = f(p^*, q^*, e) = \sup_{D \in \mathcal{D}} f(p, q, e). \tag{2}$$

For a given sample, e is fixed. Thus, the maximization takes place over the (p, q) combinations associated with each possible decision rule D in $\mathcal{D}$.

As usual, assume we observe a sample of size n from $\Pi = \langle \Pi_1, \Pi_2 \rangle$, with each observation described by k variables $X_1, \ldots, X_k$ where each variable X_i may take on s_i values, $x_{i1}, \ldots, x_{is_i}$.

Then, the General Sequential Algorithm can be described as follows:

1) Choose the variable value x_{ij} or its complement $\bar{x}_{ij}$ to maximize $\hat{O}$. That is, assume $x_{m,n} = x^{(1)}$ is such that

$$\hat{O}^{(1)}_{x_{m,n}} = \max_{\{x_{ij}, \bar{x}_{ij}\}} \hat{O}_{x_{ij}} \qquad (3)$$

That is $\hat{O}^{(1)}_{x_{m,n}}$

is the maximum value of $\hat{O}$ when evaluated for all $2 \cdot \sum_{i=1}^{k} s_i$ values that can occur. The decision rule

$$D^{(1)} = \langle D_1^{(1)}, D_2^{(1)} \rangle$$

is $D_1^{(1)} = \left\{\vec{X} \mid X_m = x_{m,n}\right\}$

$D_2^{(1)} = \left\{\vec{X} \mid X_m \neq x_{m,n}\right\}$

2) The next variable to enter is that variable value x_{ef} which maximizes $\hat{O}$ over all possible combinations $x_{m,n}$ or x_{ef}, $x_{m,n}$ and x_{ef}, $x_{m,n}$ and not x_{ef}, x_{mn} or not x_{ef}. Define the resulting composite variable as $x^{(2)}$; for example, if the maximum occurs for $x_{m,n}$ and not x_{ef}, then

$$x^{(2)} = \begin{cases} 1 \text{ if } X_m = x_n \text{ and } X_e \neq x_{ef} \\ 0 \text{ otherwise} \end{cases} \qquad (4)$$

and $D^{(2)} = \langle D_1^{(2)}, D_2^{(2)} \rangle$ with

$D_1^{(2)} = \left\{\vec{x} \mid x^{(2)} = 1\right\}$

$D_2^{(2)} = \left\{\vec{x} \mid x^{(2)} = 0\right\}$

3) The procedure continues in this fashion until a stopping rule halts the process at stage s. Some possible stopping rules are discussed later.

5.2.1 Features of the General Sequential Algorithm

Although sequential algorithms are not new (for example, the AID algorithm is sequential), the General Sequential Algorithm presented here has certain interesting features:

- The algorithm searches for "good" (in terms of the pre-specified objective function) variable values rather than variables. This is a desirable feature since we are dealing with qualitative variables.

- The algorithm is powerful in that it allows the maximization to take place over all four possible logical combinations. The AID algorithm, for example, considers only the and combination.
- The procedure makes no distributional assumptions.
- The procedure automatically considers combinations of values, although in a sequential fashion.
- The procedure is not too sensitive to the first variable selected since it allows the solution to shift away from the original variable value via the or operator.

5.2.2 Properties of the General Sequential Algorithm

More specifically, the algorithm satisfies the properties of Chapter 3:

Property 1: The procedure is clearly not dependent on the order of categories of each variable since, at each stage, all possible variable values are considered.

Property 2: By construction, the procedure handles any of the objective functions discussed in Chapter 2, including those with constraints (see 5.6).

Property 3: The form of the output is a sequence of characteristics linked by logical operators. To allocate a new case, the user merely matches the case's characteristics to this "profile." No mathematical computations are required once the profile has been constructed.

Property 4: The procedure can be used with any sample size but, unlike most other techniques, it may be most effective with small sample sizes because of the computational effort involved in the maximization at each stage.

Property 5: The procedure handles interaction effects by seeking effective combinations of values. However, the sequential approach significantly reduces the interactions to be studied.

5.2.3 Computational considerations

Despite the flexibility inherent in the General Sequential Algorithm, the computational burden does not appear too great. At any stage h, the objective function $\hat{O}^{(h)}$ must be

computed for <u>at most</u> $4 \sum_{i=1}^{k} s_i$

logical combinations (since some combinations may not be feasible, e.g., x_{ij} <u>and not</u> x_{ij}). As we have seen, the objective function is generally a simple function of P, Q, E (p, q and e).

However, from a programming standpoint the problem is more difficult. Consider just the parameters P and Q. Let $x^{(h)}$ be the composite variable defined at stage h. Now we examine all variable values to maximize O_{h+1} over all possible combinations. For variable value x_{ij} and combinations <u>or</u>, we have

$$p = \hat{P}\left(x^{(h)} \text{ } \underline{\text{or}} \text{ } x_{ij}\right)$$
$$= \hat{P}\left(x^{(h)}\right) + \hat{P}\left(x_{ij}\right) - \hat{P}\left(x^{(h)} \cap x_{ij}\right) \qquad (1)$$

$\hat{P}(x^{(h)})$ is known from computations at stage h while $\hat{P}(x_{ij})$ is known from the original data. For $\hat{P}(x^{(h)} \cap x_{ij})$, the computer must scan through all cases and count those cases for which the composite variable $x^{(h)} = 1$ <u>and</u> $x_{ij} = 1$.

Similarly,

$$q = \hat{P}\left(\pi_1 | x^{(h)} \text{ } \underline{\text{or}} \text{ } x_{ij}\right)$$
$$= \frac{\hat{P}\left(\pi_1 \text{ and } [x^{(h)} \text{ } \underline{\text{or}} \text{ } x_{ij}]\right)}{\hat{P}\left(x^{(h)} \text{ } \underline{\text{or}} \text{ } x_{ij}\right)} \qquad (2)$$

The denominator of (2) is available from (1). The numerator is found by counting those cases from π_1 which have characteristics $x^{(h)}$ <u>or</u> x_{ij}.

In effect, the sample is partitioned at each stage into those cases for which the composite variable $x^{(h)} = 1$ and those where it is zero. Thus, the computer must partition the data base and provide counts for combinations of interest. The programming for this effort turns out to be very complex.

5.2.4 Stopping rules

In general, the sequential procedure would continue through stage s, say, where s is determined by a rule such as the following:

1) $\hat{O}^{(s+1)} < O^* \leq \hat{O}^{(s)}$, where O^* is a pre-specified target value of the objective function.

2) The maximum number of variable values to be allowed in the profile is s.

3) The marginal improvement offered by the next variable is below a pre-specified value, c, say:

$$\frac{\hat{O}^{(s)} - \hat{O}^{(s-1)}}{\hat{O}^{(s-1)}} < c.$$

A number of variations on the above rules are possible. Thus, the user may specify the stopping rules appropriate for the problem under consideration.

In the next sections, we consider special cases of the General Sequential Algorithm.

5.3 Sequential Algorithm for ∇

Recall that the proportionate reduction of error measure was defined to be

$$\nabla = \frac{2P(Q-E)}{E+P(1-2E)} = f(P,Q,E)$$

For the sample, the objective function to be maximized is

$$\hat{\nabla} = \frac{2p(q-e)}{e+p(1-2e)}$$

To understand the sequential procedure, consider Table 5.1 below which shows the sample counts for each cell

TABLE 5.1

	$x^{(h)} = 1$		$x^{(h)} = 0$	
	$x_{kc}=1$	$x_{kc}=0$	$x_{kc}=1$	$x_{kc}=0$
Π_1	①	②	③	④
Π_2	⑤	⑥	⑦	⑧

where $x^{(h)}$ is the composite variable defined at stage h, and x_{kc} is another variable value being considered for combination with $x^{(h)}$ at stage h + 1.

According to the General Sequential Algorithm, four combinations may be considered. For each combination, we count the errors in terms of the error cells (cells where classification errors are made):

		Error Cells	
(i)	$x^{(h)}=1$ and $x_{kc}=1$	⑤ + ② + ③ + ④	(1)
(ii)	$x^{(h)}=1$ and $x_{kc}=0$	⑥ + ① + ③ + ④	(2)
(iii)	$x^{(h)}=1$ or $x_{kc}=1$	⑤ + ⑥ + ⑦ + ④	(3)
(iv)	$x^{(h)}=1$ or $x_{kc}=0$	⑤ + ⑥ + ③ + ⑧	(4)

For each combination, there is a corresponding value of ∇. That is, from (1),

$$\hat{\nabla}_{(i)} = \frac{2(① + ⑤)\left(\frac{①}{① + ⑤} - e\right)}{e + (① + ⑤)(1-2e)} \quad (5)$$

$$\hat{\nabla}_{(ii)} = \frac{2(② + ⑥)\left(\frac{②}{② + ⑥} - e\right)}{e + (② + ⑥)(1-2e)} \quad (6)$$

$$\hat{\nabla}_{(iii)} = \frac{2(① + ② + ③ + ⑤ + ⑥ + ⑦)\left(\frac{① + ② + ③}{① + ② + ③ + ⑤ + ⑥ + ⑦} - e\right)}{e + (① + ② + ③ + ⑤ + ⑥ + ⑦)(1-2e)} \quad (7)$$

$$\hat{\nabla}_{(iv)} = \frac{2(① + ② + ④ + ⑤ + ⑥ + ⑧)\left(\frac{① + ② + ④}{① + ② + ④ + ⑤ + ⑥ + ⑧} - e\right)}{e + (① + ② + ④ + ⑤ + ⑥ + ⑧)(1-2e)} \quad (8)$$

Thus, x_{kc} improves the prediction in terms of $\hat{\nabla}$ if

$$\max\left(\hat{\nabla}_{(i)}, \hat{\nabla}_{(ii)}, \hat{\nabla}_{(iii)}, \hat{\nabla}_{(iv)}\right) > \hat{\nabla}^{(h)} =$$

$$\frac{2(① + ② + ⑤ + ⑥)\left(\frac{① + ②}{① + ② + ⑤ + ⑥} - e\right)}{e + (① + ② + ⑤ + ⑥)(1-2e)} \quad (9)$$

Furthermore, $\hat{\nabla}^{(h+1)} > \hat{\nabla}^{(h)}$ as long as there exists at least one x_{ij} ($i=1, \ldots, k_{ij}=1, \ldots, s_i$) for which inequality (9) holds.

Expressions (5) through (8) can be written more formally in terms of the sample counts n_{ij} as follows:

$$\hat{\nabla}_{(i)} = \frac{2n_{11} - 2n_{\cdot 1}n_{1\cdot}}{n_{1\cdot} + n_{\cdot 1} - 2n_{\cdot 1}n_{1\cdot}} \tag{10}$$

$$\hat{\nabla}_{(ii)} = \frac{2n_{21} - 2n_{\cdot 2}n_{1\cdot}}{n_{1\cdot} + n_{\cdot 2} - 2n_{\cdot 2}n_{1\cdot}} \tag{11}$$

$$\hat{\nabla}_{(iii)} = \frac{2(n_{1\cdot} - n_{14}) - 2(1 - n_{\cdot 4})(n_{1\cdot})}{n_{1\cdot} + (1 - n_{\cdot 4})(1 - 2n_{1\cdot})} \tag{12}$$

$$= \frac{2(n_{1\cdot} - n_{14}) - 2(1 - n_{\cdot 4})(n_{1\cdot})}{1 + n_{1\cdot} - n_{\cdot 4} - 2(1 - n_{\cdot 4})(n_{1\cdot})} \tag{13}$$

$$\hat{\nabla}_{(iv)} = \frac{2(n_{1\cdot} - n_{13}) - 2(1 - n_{\cdot 3})(n_{1\cdot})}{1 + n_{1\cdot} - n_{\cdot 3} - 2(1 - n_{\cdot 3})(n_{1\cdot})} \tag{14}$$

$$\hat{\nabla}(h) = \frac{2(n_{11} + n_{12}) - 2(n_{\cdot 1} + n_{\cdot 2})(n_{1\cdot})}{n_{1\cdot} + n_{\cdot 1} + n_{\cdot 2} - 2(n_{\cdot 1} + n_{\cdot 2})(n_{1\cdot})} \tag{15}$$

5.4 Minimize Probability of Misclassification

As shown in Chapter II, the probability of misclassification can be written in terms of P, Q, E as

$$P(M) = E + P(1 - 2Q) \tag{1}$$

Minimizing P(M) is equivalent to maximizing the probability of correct classification,

$$P(C) = 1 - P(M) = 1 - E - P(1 - 2Q)$$

$$= f(P, Q, E) = 0_p \tag{2}$$

To further explore the nature of the sequential algorithm, we again refer to the 2 x 4 table:

	$x^{(h)}=1$		$x^{(h)}=0$	
	$x_{kc}=1$	$x_{kc}=0$	$x_{kc}=1$	$x_{kc}=0$
Π_1	①	②	③	④
Π_2	⑤	⑥	⑦	⑧

Assume $x^{(h)}$ is the composite variable defined at the hth stage and x_{kc} is a new variable value being considered at the (h+1)th stage.

For $O_p^{(h+1)} < O_p^{(h)}$ we must have, for all possible variable values x_{kc}, and for each of the four combinations, the sum of the error cell counts greater than the sum for $x^{(h)}$. Thus:

	Combinations	Error Cell Counts	
i)	$x^{(h)}=1$ and $x_{kc}=1$	⑤ + ② + ③ + ④	(3)
ii)	$x^{(h)}=1$ and $x_{kc}=0$	⑥ + ① + ③ + ④	(4)
iii)	$x^{(h)}=1$ or $x_{kc}=1$	⑤ + ⑥ + ⑦ + ④	(5)
iv)	$x^{(h)}=1$ or $x_{kc}=0$	⑤ + ⑥ + ③ + ⑧	(6)

Therefore, each of (3),(4),(5),(6) must be greater than the error cell count for $x^{(h)}$, which is ⑤ + ⑥ + ③ + ④. This implies the following must hold simultaneously:

$$\begin{aligned} ② &> ⑥ \\ ① &> ⑤ \\ ⑦ &> ③ \\ ⑧ &> ④ \end{aligned} \qquad (7)$$

That is, each of the error cells created by $x^{(h)}$ must be less than the corresponding error cell in the other row. This is possible, as shown by the following example:

	$x^{(h)}=1$		$x^{(h)}=0$		
	$x_{kc}=1$	$x_{kc}=0$	$x_{kc}=1$	$x_{kc}=0$	
Π_1	0.20	0.20	0.05	0.05	.5
Π_2	0.10	0.10	0.10	0.20	.5
	0.30	0.30	0.15	0.25	1.0

Clearly, the conditions in (7) are satisfied, and

$$\hat{O}_p^{(h)} = 1 - \hat{P}^{(h)}(M) = 1 - (0.1+0.1+0.05+0.05) = 0.7$$

$$\hat{O}^{(h+1)}_{p,x_{kc}} = \max\left(1-\hat{P}^{(h+1)}_{x_{kc}}\ (M)\right)$$

$$= 1 - \min\left((3),\ (4),\ (5),\ (6)\right)$$

$$= 1 - \min\ (.40,\ .40,\ .35,\ .45)$$

$$= 1 - 0.35 = 0.65.$$

$$\text{Thus, } \hat{O}^{(h+1)}_{p,x_{kc}} < \hat{O}^{(h)}_{p} \qquad (8)$$

This demonstrates that $\hat{O}$ is <u>not</u> always improved by the addition of another variable value. However, for the algorithm not to result in an improved $\hat{O}^{(h)}$ at each stage, condition (8) must hold for all possible variable values, i.e.,

$$\max_{\{x_{kc}\}} \hat{O}^{(h+1)}_{p,x_{kc}} < \hat{O}^{(h)}_{p} \qquad (9)$$

Given the restrictiveness of the conditions in (7) and the number of variable values available at each stage, it is highly unlikely that (9) would hold until the solution is near optimum.

The error cell counts shown in (3), (4), (5) and (6) also demonstrate that the effect of a new variable value is to shift exactly one error cell from one row to another. For example, combination i) shifts the error cell count from ⑤ + ⑥ + ③ + ④ to ⑤ + ② + ③ + ④. Hence the contribution of variable value x_{kc} in combination i) is ⑥ - ② (if the result is negative, the variable value does not help in this combination). The marginal contribution of x_{kc}, at stage h + 1 is, therefore,

$$\hat{O}^{(h+1)}_{x_{kc}} - \hat{O}^{(h)} = \max\left(② - ⑥,\ ① - ⑤,\ ⑦ - ③,\ ⑧ - ④\right) \qquad (10)$$

This can be defined more formally by writing Table 5.1 in terms of the join probabilities

$$P_{abc}, \text{ where } \begin{cases} a = 1 \text{ for } \Pi_1 \\ \quad = 2 \text{ for } \Pi_2 \\ b = 1 \text{ if } x^{(h)} = 1 \\ \quad = 0 \text{ if } x^{(h)} = 0 \\ c = 1 \text{ if } x_{kc} = 1 \\ \quad = 0 \text{ if } x_{kc} = 0 \end{cases}$$

with sample estimates $\hat{P}_{abc}$.

Then,

$$\hat{O}^{(h+1)}_{x_{kc}} - \hat{O}^{(h)} = \max\left(\hat{P}_{110}-\hat{P}_{210},\ \hat{P}_{111}-\hat{P}_{211},\ \hat{P}_{201}-\hat{P}_{101},\ \hat{P}_{200}-\hat{P}_{100}\right) \quad (11)$$

$$= \max_{c=1,2}\left(\hat{P}_{11c}-\hat{P}_{21c},\ \hat{P}_{20c}-\hat{P}_{10c}\right) \quad (12)$$

Thus, the sequential algorithm can be seen as a search, at each stage, for the variable value x_{kc} with the largest marginal contribution as given by (12). Equivalently, the search is for the variable value for which one of the four differences is a maximum.

Recall from Chapter II that minimizing the probability of misclassification is equivalent to maximizing R^2. Thus, the sequential algorithm for this situation is analogous to stepwise selection in regression analysis where the objective is to introduce the new variable that maximizes R^2, given the variables already included. The comparison is not exact, however, since in our context we introduce variable values as a logical combination with the previous composite indicator variable, rather than introducing variables to form a new regression equation.

5.5 Maximize P.Q

As discussed in Chapter II, an objective function that is a good proxy for more complicated functions is $O = f(P,Q,E) = P.Q$. That is, "good" profiles tend to have relatively high values of P and Q. Referring to Table 5.1 again, we see that

$$\hat{O}^{(h)} = \left(\frac{① + ② + ⑤ + ⑥}{n}\right)\left(\frac{① + ②}{① + ② + ⑤ + ⑥}\right) = \frac{① + ②}{n} \quad (1)$$

The corresponding values of $\hat{O}^{(h+1)}$ for the four combinations are:

		$p \cdot q$	
i)	$x^{(h)}=1$ and $x_{kc}=1$	① /n	(2)
ii)	$x^{(h)}=1$ and $x_{kc}=0$	② /n	(3)
iii)	$x^{(h)}=1$ or $x_{kc}=1$	$\frac{① + ② + ③}{n}$	(4)
iv)	$x^{(h)}=1$ or $x_{kc}=0$	$\frac{① + ② + ④}{n}$	(5)

From (4) and (5), $\hat{O}^{(h+1)} > \hat{O}^{(h)}$ unless ③ = ④ = 0. In other words, the objective function p.q can never be decreased by the addition of another variable in the or combination. In fact, the General Sequential Algorithm will force the ultimate solution to be a sequence of variable values linked by the or operation.

Furthermore, we have seen that $pq \leq e$. Yet, the available and trivial solution p = 1, q = e, is such that pq = e. No other solution could perform better in terms of this objective function.

For these reasons, pq is considered poor as an objective function if the sequential algorithm is to be applied, although the statistic pq is interesting since it reflects the expected yield of cases from Π_1. A more useful objective function is to maximize pq subject to $p \leq P^*$ where P^* is a pre-specified value. In the section that follows, we discuss the sequential algorithm when there are constraints on the objective function.

5.6 Constrained Objective Functions

In many practical problems where the aim is to identify cases in Π_1 so that they may be treated differently from cases in Π_2, resource constraints may put a limit on the value of p. In some instances, the constraint may be even more severe. For example, if the staff available for cases in Π_1 is fixed, then the value of p should not be too high (additional staff needed) or too low (staff must be laid off). Thus, there may be a target $P=P^*$, say.

When P is fixed, most of the objective functions already studied become equivalent. That is, the aim is to maximize Q subject to fixed $P = P^*$. In practice, we may wish to specify an allowable range for the realized (sample) value of p, e.g.,

$$P^* - \varepsilon < p < P^* + \varepsilon \qquad (1)$$

where, for instance, we could set

$$\varepsilon = Z_{\alpha/2} \sqrt{\frac{P^*(1 - P^*)}{n}} \qquad (2)$$

where $Z_{\alpha/2}$ is the standard normal deviate associated with a probability of $\alpha/2$ in the tail of the distribution or $\varepsilon = cP^*$, where c is a subjectively determined constant (e.g., c = 0.10).

The Generalized Sequential Algorithm could be applied to the constrained objective function so that, at each stage, the solution fell within the constraints. However, this appears to be unduly restrictive since it is only the last stage solution that needs to fall within the bounds set in (1). Thus, it may be appropriate to develop modifications to the General Sequential Algorithm.

5.6.1 Maximize 0 subject to P = P*

Algorithm 5.6.1.1

Assume that we aim to have $P^* - \varepsilon < p < P^* + \varepsilon$.

Step 1: Select x_{ij} to maximize q_{ij}. In practice, several x_{ij} may have a q_{ij} of 1. In the case of ties, select the value with the highest value of p_{ij}.

Step 2: a) If $p_{ij} < P^* - \varepsilon$, select x_{kc} to maximize

$$P\left(\Pi_1 | x_{ij}=1 \text{ or } x_{kc}=1\right) \text{ or}$$
$$P\left(\Pi_1 | x_{ij}=1 \text{ or } x_{kc}=0\right)$$

Note that $p^{(2)} > p_{ij}$. That is, we are forcing the initial value of p_{ij} higher.

b) If $p_{ij} > P^* + \varepsilon$, select x_{kc} to maximize

$$P\left(\Pi_1 | x_{ij}=1 \text{ and } x_{kc}=1\right) \text{ or}$$
$$P\left(\Pi_1 | x_{ij}=1 \text{ and } x_{kc}=0\right).$$

In this instance, we are forcing the value of p lower.

c) If $P^* - \varepsilon < p_{ij} < P^* + \varepsilon$, stop.

Step 3: Repeat step 2 starting with the composite variable $x^{(2)}$.

Although this procedure seems reasonable, it has a number of potentially critical drawbacks:

1) The procedures in step 2 do not ensure that the solution ever falls in the desired range for p. In practice, step 2b) tends to force the value of p well below $P^* - \varepsilon$ because of the and operator. Thus, the results may continuously oscillate above and below the desired range.

2) The procedure reduces the power of the General Sequential Algorithm since only two of the four combinations may be considered at any stage.

3) The procedure may reach the desired range of P* early in the process. The solution will be feasible but possibly far from optimal.

The aim, therefore, is to develop another algorithm that avoids these drawbacks. Consider the following, for instance:

- Constrain solutions at each stage so that:

$$|p^{(h+1)} - P^*| \leq |p^{(h)} - P^*|. \qquad (1)$$

  That is, at each stage h the value of $p^{(h)}$ must be closer to P*.

- Start with an unconstrained problem minimizing the probability of misclassification, or maximizing ∇, for example, using the Generalized Sequential Algorithm. Then, at (pre-specified) stage s apply the constrained sequential algorithm using the constraint of (1).

This algorithm is given below.

Algorithm 5.6.1.2

Step 1: Select x_{ij} to maximize $\hat{0}$ as in the Generalized Sequential Algorithm.

Step 2: Select x_{kc} to maximize $\hat{0}^{(2)}$ where the maximization is over all four possible combinations.

Step 3: Continue until stage s is completed. Then

a) If $P^* - \varepsilon \leq p^{(s)} \leq P^* + \varepsilon$, stop. (2)

b) If (2) does not hold, select x_{ab} to maximize $\hat{0}^{(s+1)}$ over all possible combinations with $x^{(s)}$ for which

$$|p^{(s+1)} - P^*| < |p^{(s)} - P^*|.$$

Step 4: Continue step 3 until (2) holds.

This algorithm appears to retain the maximum power of the General Sequential Algorithm while ultimately forcing the solution to the desired range. The only danger is if the solution converges slowly. This may be solved by setting an upper limit on the number of stages, s + t say (t > 0), then repeating the algorithm with a different starting variable.

We term the type of algorithm exemplified by algorithm 2 as a "delayed constraint" algorithm in the sense that the constraint (s) on the solution are not considered until the

optimization process has had a chance to work. Another modification is as follows:

Algorithm 5.6.1.3

1) Apply the Generalized Sequential Algorithm for the first s stages.

2) After stage s, continue the Generalized Sequential Algorithm until a solution occurs which satisfies the constraint, or until stage s + t, whichever comes first.

3) If a feasible solution has not occurred by stage s + t, apply the constrained algorithm to subsequent stages.

This algorithm appears to hold the greatest promise for performing consistently well, since it provides the most opportunity for the sequential approach to arrive at a good solution before the constraints are applied. At best, the unconstrained solution may happen to satisfy the constraints. At worst, the algorithm may result in an expression that involves more variable values than might be preferred.

Another type of constraint that is of interest is the form of the output. For the Generalized Sequential Algorithm the output is of the form:

$$x^{(h)} = \left(\ldots \left(\left(\left(x_1 o_2 x_2 \right) o_3 x_3 \right) o_4 x_4 \right) \ldots . o_h x_h \right)$$

where x_i denotes the variable value selected for entry at the ith stage, and

o_i denotes the logical operator (and, and not, or, or not) selected at the ith stage,

with decision rule $D = \langle D_1, D_2 \rangle$ so that

$$D_1 = \left\{ \vec{x} \mid x^{(h)} = 1 \right\}$$

$$D_2 = \left\{ \vec{x} \mid x^{(h)} = 0 \right\}$$

Note that this output format, while requiring no mathematical computation, is not easy to understand at an intuitive level because of the nesting of parenthesis. In order to be responsive to Property 3, then, we consider algorithms that are constrained to a pre-specified output form, as discussed in the next chapter.

5.7 Pre-Specified Form of Output

We have already seen that using logical operators is preferable to mathematical operators when dealing with qualitative variables. However, the output may still be complicated, as discussed above. Therefore, we consider here the development of procedures under the "constraint" of a pre-specified form of output.

5.7.1 Form 1: V or W

Let the output have the form:

$$(v_1 \text{ and } v_2 \text{ and } v_3 \text{ and } \ldots v_s) \text{ or } (w_1 \text{ and } w_2 \text{ and } \ldots \text{ and } w_t), \quad (1)$$

where the v_i and w_j are indicator variables selected by the algorithm, e.g.,

$$v_i = \begin{cases} 1 \text{ if } x_{kc}=0 \\ 0 \text{ if } x_{kc}=1 \end{cases}$$

is an indicator variable which takes the value 1 whenever x_{kc} does not occur. Then, the output of (1) implies that

$$D_1 = \left\{\vec{x} \mid (v_1=1 \text{ and } v_2=1 \text{ and } \ldots v_s=1) \text{ or } (w_1=1 \text{ and } w_2=1 \text{ and } \ldots \text{ and } w_t=1)\right\} \quad (2)$$

The expression in (1) is simplified notation for (2).

We can also rewrite (1) as

$$F = V_s \text{ or } W_t \quad (3)$$

where $V_s = (v_1 \text{ and } v_2 \text{ and } \ldots \text{ and } v_s)$

$W_t = (w_1 \text{ and } w_2 \text{ and } \ldots \text{ and } w_t)$

Thus, the output is seen to be the union of two profiles each of which is formed using the and operator. To determine whether a new case belongs to Π_1, the user needs only check whether the case has all the characteristics of V or all the characteristics of W.

Clearly, the output form could be extended to include three or more profiles, i e.,

$$F = V_s \text{ or } W_t \text{ or } Z_u \text{ or } \ldots \quad (4)$$

where $Z_u = (z_1 \text{ and } z_2 \text{ and } \ldots \text{ and } z_u)$.

At this point, there are no constraints on the component profiles V, W, Z. For example, we could have $v_i = w_j$ for some (i,j). That is, variables included in one component profile could be repeated in another component. The next question, then, is how can the profiles be developed.

Algorithm 5.7.1.1

In this approach we use sample information to develop the profile. That is, rank each individual variable value (or the absence of the variable value) with respect to the objective function of interest. Then, relabel each of the variable values in order (from best to worst) as $v_1, v_2, \ldots, v_m$,

where $$m = 2\sum_{i=1}^{k} s_i.$$

Profile V_s is developed by taking $v_1, v_2 \ldots$ until s variables are included: the value s may be predetermined or be defined such that

$$\hat{P}(v_1 \text{ and } v_2 \text{ and } \ldots v_s) \geq p^*$$

but $$\hat{P}(v_1 \text{ and } v_2 \text{ and } \ldots v_s \text{ and } v_{s+1}) < p^*,$$

where p* is a pre-specified level of p to ensure that the resulting profile does not have a trivial proportion of cases.

In practice, we may find that the joint probability rapidly approaches zero. Thus, the approach could involve the following modification, assuming a minimum p = p* again, which would help to prevent a rapid drop in the probability of the profile. Assume that the profile is to include s variables.

Step 1: Select the first variable v_i in the sequence $v_1, v_2, \ldots$, for which

$$\hat{P}(v_i) > \sqrt[s]{p^*}.$$

Enter this variable, relabeled as $v^{(1)}$.

Step 2: Select the first variable v_j in the sequence $v_1, v_2, \ldots$ for which

$$\hat{P}\left(v^{(1)} \text{ and } v_j\right) > \sqrt[s-1]{p^*}. \quad \text{Relabel } v_j \text{ as } v^{(2)}.$$

Step 3: Continue step 3 until s variables have been selected, i.e., at stage h, select the first variable v_k in the sequence $v_1, v_2, \ldots$ for which

$$\hat{P}\left(v^{(1)} \text{ and } v^{(2)} \text{ and } \ldots \text{ and } v^{(h-1)} \text{ and } v_k\right) > \sqrt[s-(h-1)]{p^*}. \tag{5}$$

At stage s, then, we have

$$\hat{P}\left(v^{(1)} \text{and } v^{(2)} \text{ and } \ldots \text{ and } v^{(s)}\right) > p^* \text{ as desired.}$$

Note: If, at any stage h, a variable cannot be found that satisfies the inequality (5), the process stops and the profile V will contain only h - 1 variables.

Now, assume the realized value of $\hat{P}(V)$ is p_v. The next step is to construct the other component, W, of the profile F = V or W. Assume that the overall target for the profile F is

$$\hat{P}(F) = p_F^*.$$

Then, the target value for the profile W is

$$p_w^* = p_F^* - p_v \tag{6}$$

since we will generally find that

$$\hat{P}(F) = \hat{P}(V) + \hat{P}(W) \text{ with } \hat{P}(V \cap W) = 0.$$

W can be constructed in the same way as V using criterion (5) on the new target value of p_w^*. However, if $p_w^* < p^*$, and the same number of variables are to be selected, profile W will be identical to V. Or, W can be constructed by restricting selection to variables not selected in construction of V. Finally, W can be constructed by repeating the process on only those observations for which $V \neq 1$. This last approach bears some resemblance to the AID algorithm in that profile V is the first "optimal" split of the data base. W represents the optimal split in the remaining portion of the data base. This process can then be repeated for further profiles Z etc... so that

$$F = V \text{ or } W \text{ or } Z \text{ or } \ldots,$$

where each component is made up of a series of variables connected by and operators.

Note that this algorithm does not follow the procedures of the General Sequential Algorithm. That is, the variables are selected in order of their unconditional effect on the objective function, rather than in a stepwise fashion based on their effect on the objective function given the variables already selected. Thus, this algorithm loses some of the power to handle interaction effects. The sequential approach is discussed next.

Algorithm 5.7.1.2

Step 1: Form the profile V exactly as with the General Sequential Algorithm, except that only the combinations and or and not are considered. Again, the variable values x_{ij} can be replaced by indicator variables v such that

$$v = \begin{cases} 1 \text{ if } X_i = x_{ij} \\ 0 \text{ if } X_i \neq x_{ij} \end{cases} \quad (7)$$

Similarly the expression not x_{ij} can be replaced by an indicator variable $\bar{v}$ such that

$$\bar{v} = \begin{cases} 1 \text{ if } X_i \neq x_{ij} \\ 0 \text{ otherwise} \end{cases} \quad (8)$$

Thus the profile V consists of variables v or $\bar{v}$ linked by the and operator, i.e.,

$$V_s = \left(v_1 \text{ and } v_2 \text{ and } v_3 \text{ and } \ldots v_s\right) \quad (9)$$

where each v_i can be expressed in form (7) or (8).

As with algorithm 1, s may be predetermined. Or, we may apply a constraint to the selection at each stage such as in (5) in order to achieve a minimum specified value of p*. Assume that the realized value of $\hat{P}(V_s)$ is p_v and that the target overall value for $\hat{P}(F)$ is p_f^*. Then, the target value for $\hat{P}(W)$ is, as before,

$$p_w^* = p_f^* - p_v.$$

Step 2: As in algorithm 1, the profile W may be constructed in several ways:

- restricting W to variables not included in V;
- developing W in the subset of observations for which $V \neq 1$.

- Developing W independently of V, but applying the restriction in (5) to ensure the target value of p_f^* is achieved.

Algorithm 5.7.1.3: Optimal Subsets

The algorithms discussed so far have involved the selection of variables one at a time. An alternative procedure is to consider optimal subsets of variables. Consider again the general profile form:

$$F = V \text{ or } W \text{ or } Z$$

where each component of the profile is made up of variables connected by the and operator, e.g.,

$$V_s = \left(v_1 \text{ and } v_2 \text{ and } \ldots \text{ and } v_s\right).$$

The aim now is to select the combination of variables (v_1, v_2, ..., v_s) to maximize the objective function of interest. In general, s would be relatively low, i.e., less than 4. For s = 2 or 3, the computational burden necessary to compute the objective function for all possible subsets is not too great, especially for small sample size. Denote the feasible subsets, ranked in order of the objective function, as S_1, S_2, ..., S_m. Then, F could be formed in the following ways:

- $V = S_1$; $W = S_2$; $Z = S_3$, etc... (10)

- Or, F could be formed by computing all possible triples S_i or S_j or S_k ($i \neq j \neq k$) and select that triple with the maximum value of the objective function, i.e., if

$$\hat{O}_{def} = \max_{i,j,k} \left(\hat{O}_{ijk} = \hat{O}\left(S_i \text{ or } S_j \text{ or } S_k\right)\right), \quad (11)$$

  set $V = S_d$, $W = S_e$, $Z = S_f$.

  In practice, the maximization could be restricted to the top x% of the S_i since it is unlikely that the optimal combination would be found among the subset S_i that perform relatively poorly with respect to the objective function.

- As usual, constraints on the resulting value of p, p*, say, may affect the approach. For example, the maximization in (11) could be restricted to

those triples for which

$$p^* - \varepsilon \leq \hat{P}\left(S_i \text{ or } S_j \text{ or } S_k\right) \leq p^* + \varepsilon, \qquad (12)$$

or the sequence in (10) could be continued until h subsets have been selected such that

$$\hat{P}\left(S_1 \text{ or } S_2 \text{ or } \ldots \text{ or } S_{h-1}\right) < p^* \text{ but}$$

$$\hat{P}\left(S_1 \text{ or } S_2 \text{ or } \ldots \text{ or } S_h\right) \geq p^*.$$

5.7.2 Form 2: V and W

In this section, we consider output of the form

$$F = V_s \text{ and } W_t, \text{ where} \qquad (13)$$

$$V_s = \left(v_1 \text{ or } v_2 \text{ or } \ldots \text{ or } v_s\right)$$

$$W_t = \left(w_1 \text{ or } w_2 \text{ or } \ldots \text{ or } w_t\right)$$

This form is like Form 1 but with the roles of or and and reversed. Thus we consider similar algorithms. Because of this similarity, details of these algorithms are not provided here

Algorithm 5.7.2.1

Select variables for V in order of their performance with respect to the objective function. If a target p_v^* is desired, the procedure should be modified so that

$$p_v^* \leq \hat{P}\ (V) \leq p_v^* + \varepsilon.$$

At each stage h, we could restrict the selection to the first variable $v^{(h)}$ such that

$$\frac{p_v^*}{h.s} \leq \hat{P}\left(v^{(1)} \text{ or } v^{(2)} \text{ or } \ldots \text{ or } \ldots\right) \leq \frac{P_v^* + \varepsilon}{h.e} \qquad (14)$$

Constraint (14) is the analog of (5) for the or combination.

W may then be constructed as in 5.7.1.1.

Algorithm 5.7.2.2

Form the profile V using the General Sequential Algorithm restricted to the combinations or and or not. Since the

AD-A081 803 MAXIMUS INC MCLEAN VA F/G 12/1
FURTHER RESEARCH INTO A NON-PARAMETRIC STATISTICAL SCREENING SY--ETC(U)
DEC 79

UNCLASSIFIED NL

2 OF 2
AD
A081803

END
DATE FILMED
4-80
DTIC

variables v_i can be used to indicate x_{kc} or not x_{kc}, the combinations are, in effect, restricted to or. Then, W can be constructed as outlined in Step 2 of Algorithm 5.7.1.1.

Algorithm 5.7.2.3

Develop V by finding the optimal (sample-based) subset of s variables linked by or. If these subsets are denoted S_1, S_2, ... in order of performance on the objective function, choose

- $V = S_1$; $W = S_2$; $Z = S_3$
- Or, find the optimal combination of S_i and S_j and S_k, i.e., find S_d, S_e and S_f such that, for $d \neq e \neq f$,

$$\hat{O}_{d,e,f} = \max_{i,j,k} \left(\hat{O}_{i,j,k_s} = \hat{O} \left(S_i \text{ and } S_j \text{ and } S_k \right) \right).$$

- Develop optimal subsets with constraints on the objective function.

5.8 Evaluation

The procedures described here have a great deal of intuitive appeal. However, the complexity of the logical expressions being developed makes it difficult to develop closed-form results that would establish statistical properties, for example, the convergence of the sequence of values $\hat{O}^{(h)}$ to the true optimal value O^*.

As discussed in 5.2.2, the General Sequential Algorithm does satisfy the five properties defined in Chapter II. Nonetheless, a test of the approach(es) would be appropriate. The steps involved in such a test are:

1) Develop a general computer program to handle the General Sequential Algorithm and its variations. This program would be set up so that the user could specify in advance the objective function desired, constraints on system parameters, the format of output, and the stopping rule.

2) Obtain a data base of observations from a population $\Pi = < \Pi_1, \Pi_2 >$ with each observation described by a set of categorical variables $X_1, \ldots, X_k$.

3) Randomly divide the data base into two halves, holding one back.

4) Apply the General Sequential Algorithm, and its variations, to the other half. For each approach used, obtain the value of the objective function $\hat{O}_B$.

5) For each profile developed in 4., calculate the value of the objective function on the hold-out sample, $\hat{O}_A$. Compare $O_A - O_B$ (where possible, use estimates of the asymptotic variance of $\hat{O}_A$ and $\hat{O}_B$ to perform statistical tests of hypothesis, e.g., the procedure for doing this for the objective function ∇ was provided in Chapter II).

6) Apply the LDF, Regression Analysis and AID techniques to the same data base. Compare the $\hat{\nabla}$ results for these procedures on the test and hold-out sample as in 5. Also, compare the $\hat{\nabla}$ results for these procedures with the $\hat{\nabla}$ results for the General Sequential Algorithm and its variations.

We have, however, some empirical evidence of the effectiveness of one approach as applied to the Medicaid data base, as shown in Table 5.2 below:

TABLE 5.2: TEST RESULTS

Type of Case	Profile Number	p	q	e	$\hat{\nabla}$	$\hat{P}$ (Misclassification)
AI	1.1	0.1	0.76	0.17	0.50	0.12
	1.2	0.2	0.61	0.17	0.58	0.12
NH	2.1	0.1	0.92	0.30	0.36	0.22
	2.2	0.2	0.77	0.30	0.44	0.19
AFDC	3.1	0.1	0.77	0.27	0.32	0.22
	3.2	0.2	0.58	0.27	0.36	0.24

These results were based on a version of the algorithm with pre-specified output form, of the type 5.7.2, with constraints on the value of p (0.1 or 0.2). Again, the intent was not to maximize $\hat{\nabla}$, but to maximize q for the fixed value of p. Thus the $\hat{\nabla}$ values tend to understate what could have been achieved. Furthermore, since the solution was highly constrained, we can

hypothesize that results might have been superior if the unconstrained General Sequential Algorithm had been used.

5.9 Summary

In this Chapter, we have introduced a new class of screening techniques for handling qualitative variables. Specifically, we have done the following:

- defined a General Sequential Algorithm that can be used with a wide range of objective functions;
- developed algorithms that can be used to achieve a pre-specified form of output.

In general, these algorithms were developed from an intuitive and practical standpoint. Hence, the choice among algorithms is, at this point, largely a function of the user's preference. For example, if the form of the output is of paramount concern, then one of the algorithms in 5.7.1 and 5.7.2 should be adopted.

The algorithms were also developed under the assumption that no human intervention would be used in the construction of profiles. In practice, however, human intervention may be used just as it is in, say, regression analysis where a large number of regression runs may be used to guide the final model specification and estimation. The flexibility of the approaches described here is such that different techniques can be applied to the same problem, changes in the number of variables to be included can be made, constraints can be relaxed, etc.... The user may then choose the "best" among several solutions using criteria above and beyond those inherent in the objective function.

Thus, we have attempted to widen substantially the range of techniques available to those interested in discriminating between two populations based on the qualitative characteristics of cases in each population.

However, the algorithms are all of the "search" type wherein the data are analyzed in depth in order to identify effective decision rules D. As noted, the rules are developed Ex Post to fit the data. Because of the flexibility of the approaches, the degrees of freedom are very high. Thus, it is very important that the methodologies only be applied in instances where it is possible to test the results on a hold back or new sample before

adopting the solution. The techniques are very powerful from the viewpoint of explaining the data, but this power may be dangerous if used carelessly.

Furthermore, the algorithms can be expected to require extensive computer time because of the number of calculations to be performed at each stage. The rapid advances in computer technology have made such approaches possible. It seems appropriate that statistical technology keep pace with the power of computers. Hopefully, the work presented here is a step in this direction.

REFERENCES

REFERENCES

Aitchison, J. and Aitken, C.G.G. [1976]. "Multivariate Binary Discrimination by the Kemel Method." Biometrika 63, No. 3, 413-20.

Baker, Kenneth and Albaum, Gerald [1976]. "The Sampling Problem in Validation of Multiple Discriminant Analysis." Journal of Market Research Society 18, No. 3, 158-61.

Bendel, Robert B. and Afifi, A.A. [1977]. "Comparison of Stopping Rules in Forward 'Stepwise' Regression." JASA 72, No. 357, 46-53.

Bishop, Y.M.M.; Fienberg, S.E. and Holland, P.W. [1975]. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Broffit, James D.; Randles, Ronald H.; and Hogg, Robert V. [1976]. "Distribution Free Partial Discriminant Analysis." JASA 71, No. 356.

Chickner, Robert P. [1976]. "On Least Squares Estimation for Categorical Data." Communications in Statistics - Theory and Methods. A5. No. 11, 1059-64.

Cochran, W.G. [1964]. "On the Performance of the Linear Discriminant Function." Technometrics 6, 179-90.

Cochran, W.G. and Hopkins, C.E. [1961]. "Some Classification Problems with Multivariate Qualitative Data." Biometrics 17, 10-32.

Costner, H. [1965]. "Criteria for Measures of Association." American Sociological Review 30, 341-353.

David, Jean M. and McIver, Carolyn. [1976]. "An Application of Multivariate Discriminant Analysis to Perceptions of Student Personnel Services." American Statistical Association. Proceedings of the Social Statistics Section, Part I, 270-275.

Dawson, Beth [1976]. "A Review of the Applicability of Discriminant Analysis to Social Science Research." American Statistical Association, Proceedings of the Social Statistics Section, Part I, 276-281.

Dillon, W.R. and Goldstein, M. [1978]. "On the Performance of Some Multinomial Classification Rules." JASA 78, No. 362.

DiPillo, P.J. [1970]. "The Application of Bias to Discriminant Analysis." Communications in Statistics, Theory and Method, A5. No. 9, 843-54.

Dunn, O.J. and Varady, P.D. [1966]. "Probabilities of Correct Classification in Discriminant Analysis." Biometrics 22, 908-24.

Gail, Mitchell M. and Green, Sylvan B. [1976]. "A Generalization of the One-Sided Two Sample Kolmogorov-Smirnov Statistic for Evaluating Diagnostic Tests." Biometrics 32, No. 3, 561-70.

Gilbert, E.S. [1969]. "On Discrimination Using Qualitative Variables." JASA 63, 1399.

Goldstein, M. and Wolf, C. [1977]. "On the Problem of Bias in Multinomial Classification." Biometrics 33, 325-331.

Goodman, Leo A. and Kruskal, William M. [1954]. "Measures of Association for Cross-Classifications." JASA 49, 732-64.

Grizzle, J.E.; Starmer, C.F. and Koch, G.G. [1969]. "Analysis of Categorical Data by Linear Models." Biometrics 25, 489-504.

Harushek, Eric A. and Jackson, John E. [1977]. Statistical Methods for Social Scientists. Academic Press, Inc., New York.

Hartigan, John A. [1975]. Clustering Algorithms. John Wiley and Sons, Inc.

Hildebrand, David K.; Laing, James D. and Rosenthal, Howard [1977]. Prediction Analysis of Cross Classifications, John Wiley and Sons, Inc.

Hills, M. [1966]. "Allocation Rules and Their Error Rates." Journal of the Royal Statistical Society, B28. 1-31.

Lachenbruch, P.A. and Mickey, M.R. [1968]. "Estimation of Error Rates in Discriminant Analysis." Technometrics 10, No. 1.

Landis, Richard J.; Freeman, Jean L.; Stanish, William M.; Koch, Gary G.; and Lewis, Alcinda L. [1976]. "GENCAT: A Computer Program for the Generalized Least Squares Analysis of Multivariate Categorical Data." American Statistical Association. Proceedings of the Statistical Computing Section, 190-195.

Lehmann, E.L. [1959]. Testing Statistical Hypotheses. New York, John Wiley and Sons, Inc.

Martin, D.C. and Bradley, R.A. [1972]. "Probability Models Estimation and Classification for Multivariate Dichotomous Populations." Biometrics 28, 203-222.

Matsuita, K. [1954]. "On Estimation by the Minimum Distance Method." Am. Inst. Stat. Math. 7, 67-77.

- [1955]. "Decision Rules Based on the Distance for Problems of Fit, Two Samples and Estimation." Am. Math Stat. 26, 631-40.

- [1957]. "Classification Based on Distance in Multivariate Gaussian Cases." Proc. Fifth Berkeley Symp. Math. Stat. and Prob., 1, 299-304.

McDonald, Lyman L.; Lowe, Victor W.; Smidt, Robert K.; Meister, Keven A. [1976]. "A Preliminary Test for Discriminant Analysis Based on Small Samples." Biometrics 32, No. 2. 417-22.

McLachlan, G.J. [1974]. "Estimation of the Errors of Misclassification on the Criterion of Asymplotic Mean Square Error." Technometrics 16, 256-60.

McLachlan, G.J. [1976]. "The Bias of the Apparent Error Rate in Discriminant Analysis." Biometrika 63, 239-244.

McLachlan, G.J. [1976]. "A Criterion for Selecting Variables for the Linear Discriminant Function." Biometrics 32, No. 3, 529-34.

Miller, R.G. [1974]. "The Jackknife - A Review." Biometrika 61, 1-15.

Moore, D.H. [1973]. "Evaluation of Five Discriminant Procedures for Binary Variables." JASA 68, 399-404.

Mosteller, F. [1968]. "Association and Estimation in Contingency Tables." JASA 63, 1-28.

New Hampshire Division of Welfare [1977]. First Year Report on the Title XIX Quality Control Project, Project #11-P-90147. Office of Research and Demonstrations, DHEW.

- [1978]. Second Year Report on the Title XIX Quality Control Project.

- [1979]. Third Year Report on the Title XIX Quality Control Project.

Oberhue, H. and Ono, M. [1976]. "A Statistical Micro-Data Source on AFDC Recipients." American Statistical Association. Proceedings of the Social Statistics Section, Part II, 645-50.

Powers, John A.; March, Lawrence, C.; Huckfeldt, Robert R.; Johnson, C.L. [1978]. "A Comparison of Logit, Probit and Discriminant Analysis in Predicting Family Size." American Statistical Association, Proceedings of the Social Statistics Section, 693-698.

Quesenberry, and Gessaman [1968]. "Nonparametric Discrimination Using Tolerance Regions." Annals of Math. Stat., April, 664-73.

Reinmuth, James E. and Hawkins, Del I. [1977]. "Qualitative Variable Discriminant Analysis and Its Use in Product Version Selection." Decision Sciences 8, No. 2. 478-88.

Ralph, John E.; Williams, Albert P. and Lee, Carolyn L. [1978]. "The Effect of State of Residence on Medical School Admissions: Empirical Bayes and Least Squares Discriminant Estimators." American Statistical Association, Proceedings of the Social Statistics Section, Part I, 89-98.

Scott, A.J. and Knott, M. [1976]. "An Approximate Test for Use with AID." Applied Statistics 25, No. 2, 103-106.

Social Security Administration, Office of Family Assistance. [1979]. Conference on the Utilization of Characteristic Profiles as a Workload Planning Technique. Conference Notebook.

Sonquist, John A.; Baker, Elizabeth, L. and Morgan, James N. [1973]. Searching for Structure. University of Michigan, Ann Arbor, Michigan.

Sorum, M. [1971]. "Expected and Optimal Probabilities of Misclassification." IEEE Trans. on Information Theory, IT-20, 472-479.

Welch, B.L. [1939]. "Note on Discriminant Functions." Biometrika 31, 218-20.